Title: What are the variables associated with the interest rate of a loan? Identification and quantification associated variables.

Introduction:

Who didn't apply for a loan from a bank? How much data did you have to provide to the bank about you? ,.. Imagine for a moment that you want to be a banker or you want to lend money to people. The first question could be: How will I determine the interest rate of a loan for an applicant? , the second question could be: What are the variables associated with the interest of a loan?, and the third question could be: How can we identify and quantify the variables associated with the interest of a loan?

In this analysis, the interest rate is our response variable, and the result of this variable depends on one or more independent variables. In our case, we have several explanatory variables about each applicant such as the amount of money requested in the loan application, the amount of money loaned to the individual, the term of the loan (time), the number of open lines of credit the applicant had at the time of application, the duration of employed time at current job, the monthly income of the applicant, the measure of the creditworthiness of the applicant,…

The interest rate of a loan depends on the level of the applicant´s risk, that is if an applicant has good monthly income, a good measure of creditworthiness,… then this applicant will have less risk of non-payment its loan and therefore the interest rate will be lower than an applicant has more risk of non-payment.

So, in this analysis we performed a study to understand the associations or the relationships of the various available variables in the data set that can help us to determine the interest of loan. We used exploratory analysis to understand better the data of data set and get intimately familiar with them. We performed basic data summaries and visualization (by plotting), and we could detect trends, patterns, anomalies and outliers. Using multiple regression methods we obtained a regression model that could us there are significant relationship between the interest rate of loan and several associated variables such as the amount (in dollars) requested in the loan application, the amount (in dollars) loaned to the individual, the percentage of consumer's gross income that goes toward paying debts [1], the number of open lines of the credit of the applicant, the total amount outstanding all lines of credit [2], the number of authorized queries about applicant creditworthiness in the term of 6 months [3], the term of loan (in months) and the FICO score[4] (the measure of the creditworthiness of the applicant).

Methods:

Data Collection

For this analysis we used a data set on peer-to-peer loans issued through the platform Lending Club [5]. This data set contains information of people applying for a loan such as the duration of employed time at current job, the measure of their creditworthiness (FICO Score), their monthly incomes,… which are used to determine the associated interest rate to a loan. The data were downloaded from coursera.org [6] in Data Analysis Course on February 13, 2013 using the R programming language.

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. We identified transformations to perform on the raw data on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, (3) perform transformations of the skewed distributions and (4) determine the terms used in the regression model relating the interest rate and other associated variables.

Statistical Modeling

To identify and quantify the association between the interest rate of a loan and other variables, we performed a standard multivariate linear regression model [7]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between the amount (in dollars) requested in the loan application, the amount (in dollars) loaned to the individual, the percentage of consumer's gross income that goes toward paying debts, the number of open lines of the credit the applicant, the total amount outstanding all lines of credit, the number of authorized queries about applicant creditworthiness in the term of 6 months, the term of loan (in months) and the FICO score (the measure of the creditworthiness of the applicant). Coefficients were estimated with ordinary least squares and standard errors were calculated using standard asymptotic approximations [8][9].

Reproducibility

All analyses performed in this manuscript are reproduced in the R markdown file loansLendingClubFinal.Rmd [10].

Note. Due to security concerns with the exchange of R code, we don't submit code to reproduce analysis, in this data analysis,

Results:

The sample of loans data used in this analysis is a total size 2500 observations with 14 variables that contains the following information, the amount of money (in dollars) requested in the loan application (Amount.Requested), the amount of money (in dollars) loaned to the individual (Amount.Funded.By.Investors), the lending interest rate (Interest.Rate), the length of time (in months) of the loan (Loan.Length), the purpose of the load (Loan.Purpose), the percentage of consumer's gross

income that goes toward paying debts [1] (Debt.To.Income.Ratio), the abbreviation for the U.S. state of residence of the loan applicant [11] (State), a variable indicating whether the applicant owns, rents, or has a mortgage on their home (Home.Ownership), the monthly income (in dollars) of the applicant, a range indicating the applicants FICO score [4] (FICO.Range), the number of open lines of credit the applicant had at the time of application (Open.CREDIT.Lines), the total amount outstanding all lines of credit [2] (Revolving.CREDIT.Balance), the number of inquires about the creditworthiness of the applicant in the term of 6 months before the loan was issued [3] (Inquiries.in.the.Last.6.Months) and the duration of employed time at current job (Employment.Length). We eliminated missing values in the dataset, specifically there were seven missing values (NA) that affected only two rows. There were also incorrect data in observations related to the amount (in dollars) loaned (Amount.Funded.By.Investors) with values of 0.0 and -1.0, we corrected these values with the same value as the amount (in dollars) requested (Amount.Requested). In the duration of employed time (Employment.Length), we found rows in the data set with "n/a" values and here we didn't do anything on them. Too, due to detected outliers for the monthly income and the total amount outstanding of all line credit variables, we removed involved rows.

The distributions of the amount requested, the amount loaned, the monthly income, the number of open lines of credit and the total amount outstanding all lines of credit were **right-skewed**, we performed transformations (log10) on these variables to improve their normality. In the case of the number of authorized queries in the term of 6 months we applied on a transformation of squared root (sqrt) type, because it is a count variable.

After we performed a lot regression models for this dataset, our final regression model was:

$$\begin{aligned} \text{Interest.Rate} = b_0 &+ b_1 \log10(\text{Amount.Requested}) + b_2 \log10(\text{Amount.Funded.By.Investors}) \\ &+ b_3 \text{Debt.To.Income.Ratio} + b_4 \log10(\text{Open.CREDIT.Lines}) \\ &+ b_5 \log10(\text{Revolving.CREDIT.Balance}) + b_6 \sqrt{\text{Inquiries.in.the.Last.6.Months}} \\ &+ f(\text{Loan.Length}) + g(\text{FICO.Range}) + e \end{aligned}$$

Term	Description
b_0	Intercept
b_1	The change in the interest rate of loan with a change of 1 unit in log base 10 dollars for the amount requested in the loan
b_2	The change in the interest rate of loan with a change of 1 unit in log base 10 dollars for the amount loaned to the individual
b_3	The change in the interest rate of loan with a change of 1 unit in log base 10 percent for the percentage of consumer's gross income
b_4	The change in the interest rate of loan with a change of 1 unit in log base 10 units for the number open lines
b_5	The change in the interest rate of loan with a change of 1 unit in log base dollars for the total amount outstanding all lines of credit
b_6	The change in the interest rate of loan with a change of 1 unit in squared root units for the

	number of authorized queries
f(Loan.Length)	Represent factor models (the length of time of the loan) with 2 levels (36months and 60 months)
g(FICO.Range)	Represent factor models (a range indicating the applicants FICO Score) with 36 levels (640-644, 645-649, ……. 820-824 and 830-834)
e	Term error. Represents all sources of unmeasured and unmodeled random variation in interest rate of loan

The result of our multiple linear regression model offers us, the linear relationship between the interest rate and all previous terms that are highly statistically significant (p-value= **2.2e-16**) and how well the regression model fits the observed data is **0.7942**. We observed a strong association between the interest rate and FICO score.

Conclusions:

Clearly, in this analysis, there is a strong association between the interest rate of a loan and the FICO Score Range. To fit still better our regression model we included other variables such as the amount of money requested, the amount of money loaned, the DTI (Debt To Income Ratio), the number of open lines of credit the applicant, the total amount outstanding all lines of credit, the number of authorized queries and the length of time employed. We included these last variables in the our regression model, to get a high R-squared percentage (around 80%) and highly statistically significant, in spite of what some included variables have association with the interest rate less than the FICO Score Range.

References

[1] Debt-to-Income ratio
http://en.wikipedia.org/wiki/Debt-to-income_ratio. Accessed 02/17/2013
[2] Revolving Credit Balance
http://www.ehow.com/about_7550001_revolving-credit-balance.html. Accessed 02/17/2013
[3] Inquiries in the last 6 months
http://www.myfico.com/crediteducation/creditinquiries.aspx. Accessed 02/17/2013
[4] FICO Score
http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. Accessed 02/17/2013
[5] Lending Club
https://www.lendingclub.com/. Accessed 02/17/2013
[6] Data set of loans
https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. Accessed 02/17/2013
[7] Jank, Wolfgang. *Business Analytics for Managers.* Springer, 2011
[8] Ferguson, Thomas S. *A Course in Large Sample Theory: Texts in Statistical Science.* Vol.l 38. Chapman & Hall/CRC, 1996.

[9] Shahbaba, Babak. *Biostatistics with R. An Introduction to Statistics Through Biological Data.* Springer, 2012
[10] R Markdown Page.
http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 02/13/2013
[11] The abbreviation for U.S. state
http://www.50states.com/abbreviations.htm#.USHpnWd1z2R. Accessed 02/17/2013