Title: Increased interest rate is associated with increased loan length period

Introduction:

In this study we show the association between interest rate of loans and the other variables in the Lending Club data set. In particular, we will consider whether any of these variables have an important association with interest rate after taking into account the applicant's FICO score [1]. If FICO score is the only variable affecting the loan interest rate, one will find a straight line on a plot showing the relation between them. Yet this is not the case; i.e. there are variations on the interest rates for applicants having the same FICO score. Our study attempts to identify which other variables might be affecting the interest rates.

Methods:

Data Collection

The data for our analysis consist of a sample of 2,500 peer-to-peer loans issued through the Lending Club [2]. The interest rate of these loans is determined by the Lending Club on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores. The data were downloaded from the Coursera Data Analysis web page on February 17, 2013 using the Python programming language [3].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms used in the regression model relating interest rate to other variables. We highlight specifically the loan length, as it appears to be the one showing most correlation with the interest rate.

Statistical Modeling

To relate interest rate to interest rate we performed a standard multivariate linear regression model [4]. Model selection was performed mainly on the basis of our exploratory analysis. Coefficients were estimated with ordinary least squares and standard errors were calculated using standard asymptotic approximations [5].

Reproducibility

All analyses performed in this manuscript are reproduced in Python programming language, developed on IPython notebook environment[1] [4].

Results:

The data used in this analysis contains information on the:

- amount (in dollars) requested in the loan application,
- amount (in dollars) loaned to the individual,
- lending interest rate,
- loan length of time (in months),
- purpose of the loan as stated by the applicant,
- debt-to-income ratio of the applicants, that is, the percentage of the applicant's gross income that goes toward paying debts,
- state where the applicant resides,
- home ownership status of the applicant,
- applicant's monthly income (in dollars),
- applicant's FICO range, measuring his/her creditworthiness,
- number of open lines of credit the applicant had at the time of application,
- applicant's total amount outstanding all lines of credit,
- applicant's total number of creditworthiness inquiries in the 6 months before the loan was issued, and
- applicant's length of time employed at the current job.

Of the 2,500 observations, we identified two that contain missing values, which were simply removed, because we expect that such few occurrences will not affect the analysis results significantly.

The interest rates are evenly distributed, with near-bell-shape distribution centered at around 13%. Based on this observation we identified that no transformation is required on the interest rate.

We also split the variable types into numerical and categorical. Most of the time this is obvious from the data. However it is not the case for the loan length. We choose to treat it as categorical as there are only two (numerical) values in the data, 36 and 60 months, and we believe it may not make much sense if our model is eventually used for values other than these two numbers.

We first check the correlation between all numerical variables in the data. We exclude the amount of granted loan from further analysis because it is highly correlated with the amount of requested loan, which, common sense says, might affect the interest rate of the loan. Our

[1] The notebook can be accessed on http://nbviewer.ipython.org/4974422/ either for viewing online or to reproduce the exact results presented in this manuscript.

correlation table shows that it is indeed the case. We further identified other variables that could be useful for further analysis based on their relative magnitudes of correlation with respect to the interest rate.

Some exploratory analyses were needed to determine the significance of the categorical variables. We identified that the loan length has a stronger effect on the interest rate compared to other categorical variables. We can verify this by looking at boxplots of the interest rate grouped by the loan lengths. A scatter plot between the FICO range and interest rate, grouped by loan lengths, also shows some clustering, although there are plenty of overlaps. This is shown in the attached figure.

We fit a regression model relating interest rate to loan length and FICO range. Although we could still improve the fit by including the numerical variables, we choose not to do it to make a simple and tractable model. We also don’t consider interaction between loan length and FICO range. Our final regression model was:

$$R = b_0 + \ b_1 L + \sum_{i=1}^{N_F - 1} c_i F_i + e$$

where b_0 is the interest rate for loan length time of 36 months and FICO range 640-644 (the first range in the group). *L* is the dummy regressor variable for loan length, and F_i’s are the dummy variables for the FICO ranges. The error term *e* represents all sources of unmeasured and unmodelled random variation in interest rate. Our final regression model appeared to contain no non-random patterns of variation in the residuals.

We observed a highly statistically significant (*P* = 1.56e-282) association between interest rate and loan length. A 60-month loan correspond to a higher interest rate of b_1 = 4.38% of interest rate (95% confidence interval: 4.17, 4.59). So for example, if two people within the same FICO score range (all other numerical factors considered equal), one of them applies for a certain loan for 36 month and received an interest rate of 15%, the other is expected to have to repay with an interest rate of 19.38%.

Conclusions:

Our analysis suggests that there is a significant, positive association between interest rate and loan length. We also observe strong correlation of the numerical variables, which could be included in the model to increase the fit and prediction accuracy. However we do not perform this exercise in order to keep the model simple and tractable.

References

[1] Wikipedia credit score in the United States page. URL: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. Accessed: 17.02.2013.

[2] The Lending Club page. URL: https://www.lendingclub.com/home.action. Accessed 17.02.2013.

[3] Python programming language. URL: http://www.python.org. Accessed: 17.02.2013

[4] IPython interactive computing. URL: http://ipython.org. Accessed: 17.02.2013

[5] Ferguson, Thomas S. *A Course in Large Sample Theory: Texts in Statistical Science*. Vol. 38. Chapman & Hall/CRC, 1996.