Title: Analyzing factors correlated with increased interest rates in Peer-to-Peer loans

Introduction:

LendingClub.com is a corporation that facilitates peer-to-peer loans for up to $35,000. Peer-to-peer loans are loans that are funded by an individual or group of individuals rather than by a financial institution. LendingClub.com has existed since 2007 and boasts over $2 Billion in loans made. [1]

Interest rates for loans are determined by a number of factors that determine repayment risk and the factors used vary by lending organization. [2] The purpose of this study is to examine the factors that are most closely associate with the resulting interest rate on the loans granted by LendingTree.com.

Methods:

Data Collection

For our analysis, data was used from LendingClub.com. This data is a de-identified random sample of 2500 issued peer-to-peer loans.

Exploratory Analysis

Exploratory Analysis was performed by examining plots, tables and summary statistics on the observed data. A number of transformations were applied to the raw data to make data more useful for analysis, including data typing and scaling. Missing values were identified, data quality was assessed, and preliminary analysis performed to determine the terms for use in the regression model relating various details about the loan to the resulting interest rate.

Statistical Modeling

To relate interest rate to the other details of the loan, we performed standard multi-variate linear regression, using accepted models. [3] Model selection was performed on the basis of knowledge gained through exploratory analysis, along with general knowledge of the nature of interest rate determination in lending. Coefficients were estimated using standard least squares and error measures calculated using standard approximations. [4] [5]

Reproducibility

All analysis performed for this assignment are reproduced using standard R code. To reproduce this analysis exactly, the source file provided by LendingClub.com must also be acquired.

Results:

The data used in this analysis contains information about each of the 2500 loans that were issued. This data includes (AR) loan amount requested, (AF) amount funded by investors, (IR) interest rate applied to the loan, (LL) the length of the loan, (LP) the purpose of the loan, (DI) the applicants debt-to-income ratio, (ST) the state in which the loan was issued, (CB) revolving credit balance, (HO) the applicant’s home ownership status, (MI) the applicant’s monthly income, (FI) the applicant’s FICO score, (CL) the applicant’s open credit lines, (IQ) the number of the applicant’s credit inquiries in the past 6 months, and (EL) the length of employment.

A number of missing values were found in the data set in the , (CB) revolving credit balance, (CL) the applicant's open credit lines and, (MI) the applicant's monthly income fields. The total number of records affected was 2. These records were eliminated from the analysis.

Several fields were heavily rightly skewed. These fields were (CB) revolving credit balance, (CL) the applicant's open credit lines and, (MI) the applicant's monthly income. To aid in linear regression techniques, a log base 10 transform was applied to these variables. Subsequent analysis used these transformed variables.

A regression model was first fit relating interest rate to FICO score. Residuals showed non-random variation, leading to the analysis of potential confounders in the data. Additional models were fitted based on analysis. Each of the variables applied is depicted as a panel in Figure 1. The final regression model used was:

$$IR = b_0 + b_1(FI) + f(LL) + g\big(log10(CB)\big) + h(AF) + e$$

Where b_0 is the intercept term and $b_1$represents the coefficient of the factored FICO score associated with the loan. The term f(LL) is a factored term corresponding to the two loan length options of "36 months" and "60 months". $g\big(log10(CB)\big)$ and $h(AF)$ represent continuous variables. The error term e represents all random variation in the model not explained by the factors included in the model.

We observe a highly statistically significant (P < .0001) association with the variables fitted to the linear model and the interest rate given on the loan. A change of one unit of the FICO score factor corresponded to a -.45 unit change in the interest rate (95% confidence interval: -.46, -.43). A change of one unit in the factored (LL) loan length variable corresponded to a 3.2 unit change in the interest rate (95% confidence interval: 3.02, 3.45). A change of one unit log base 10 in (CB) revolving credit balance corresponded to a -.26 unit change in the interest rate (95% confidence interval: -.30, -.20). A 1 unit change in the (AF) amount of the loan corresponded to a 1.55e-04 unit change in interest rate (95% confidence interval: 1.4e-04, 1.67e-04).

Conclusions:

Our analysis suggests that there is a significant association between interest rate and the following variables: FICO score, Loan Length, Revolving Credit Balance, and Loan Amount. While the strongest association is between interest rate and FICO score, the addition of the above included variables improves the fitment of the model without reducing the correlation.

This analysis is limited to only a small subset (2500) of a specific type of loan (peer-to-peer). A larger data set would produce more robust results. Further, this model should not be used to understand the relationship between these variables and the interest rate on other types of loans; this should be a subject of future research with a more comprehensive data set.

References:

[1] "About US," 16 11 2013. [Online]. Available: https://www.lendingclub.com/public/about-us.action.

[2] "Interest Rates," Wikipedia.org, 15 11 2013. [Online]. Available: http://en.wikipedia.org/wiki/Interest_rate#Risk. [Accessed 15 11 2013].

[3] G. A. a. A. J. L. Seber, "Linear Regression Analysis," *Wiley,* vol. 936, 2012.

[4] T. S. Ferguson, "A Course in Large Sample Theory: Texts in Statistical Science.," *Chapman & Hall/CRC,* vol. 38, 1996.

[5] J. Leek, "Sample Analysis Assignment," Coursera.com, 2013.