Name: José Carlos Baquero Triguero
Instructor's Name: Jeff Leek
Course: Data Analysis
Date: February 18, 2013

Analysis of interest rate on loans

Introduction

An interest rate is the rate at which interest is paid by borrowers for the use of money that they borrow from a lender. Specifically, the interest rate (I/m) is a percent of principal (P) paid at some rate (m) [1]. The interest rate is determined on the basic of characteristics of the person asking for the loan such as their employment history, credit history and creditworthiness score like FICO Score[2].

Understanding the relationship of interest rate on loans can help us characterize the loans and lenders. Here we performed an analysis to determine if there was a significant association between interest rate and characteristics of the borrowers. Using exploratory analysis and standard multiple regression techniques we show that there is significant relationship between Interest Rate and FICO Range, even after adjusting for important confounders such as Amount Requested and Amount Founded by Investor. Our analysis suggests that loans with high FICO Range have less interest rate.

Methods:

Data Collection:

For our analysis we use the data consist of a sample of 2,500 peer-to-peer loans issued through the Lending Club (https://www.lendingclub.com/home.action). The data were downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv on February 14, 2013

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, and (3) determine the terms used in the regression model relating interest rate of loans and the others variables in the data set.

Statistical Modeling:

To relate interest rate to characteristic of borrowers we performed a standard multivariate linear regression model [4]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between interest rate and borrower. 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 the R markdown file interestRateFinal.Rmd [3]. To reproduce the exact results presented in this manuscript the cached version of the analysis must be performed, as the data available from https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

Results:

The loans data used in this analysis contains next information:

- Amount Requested - The amount (in dollars) requested in the loan application
- Amount Funded By Investors – The amount (in dollars) loaned to the individual
- Interest rate – The lending interest rate.
- Loan length – The length of time (in months) of the loan
- Loan Purpose – The purpose of the loan as stated by the applicant
- Debt.to.Income.Ratio – The percentage of consumer's gross income that goes toward paying debts [6]
- State – The abbreviation for the U.S. state of residence of the loan applicant [7]
- Home Ownsership - A variable indicating whether the applicant owns, rents, or has a mortgage on their home.
- Monthly income – The monthly income of the applicant (in dollars).
- FICO range – A range indicating the applicants FICO score. This is a measure of the creditworthiness of the applicant. [2]
- Open Credit Lines – The number of open lines of credit the applicant had at the time of application.
- Revolving Credit Balance – The total amount outstanding all lines of credit [8]
- Inquiries in the Last 6 Months – When a person applies for credit, they authorize the lender to "inquire" about their creditworthiness [9]
- Employment Length – Length of time employed at current job.

We identified seven missing values in the data set we collected and also all measured variables were observed to be inside the standard ranges.

Most borrowers had standard monthly incomes (less than $6.800/month - 75% of borrowers). The distribution of monthly incomes was heavily left skewed. Based on the distribution of the monthly incomes we recognized that a transform was necessary to improve the performance of linear regression techniques; we performed a log base 10 transform of the monthly incomes. Subsequent analyses focus on this transformed monthly income variable.

Also most borrowers had normal revolving credit balance (less than %18.889 - 75% of borrowers). The distribution of revolving credit balance was heavily left skewed. Based on the distribution of the revolving credit balance we recognized that a transform was necessary to improve the performance of linear regression techniques; we performed a log base 10 transform of the revolving credit balance. Subsequent analyses focus on this transformed revolving credit balance variable.

We first fit a regression model relating Interest Rate to FICO Range. The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including potential confounders. Our final regression model was:

$$IR = b0 + b1(FICO) + f(AR) + g(AFI) + e$$

Where b0 is an intercept term and b1 represents the change in FICO Range on the Interest Rate associated with a change of 1 unit. The terms f(AR) and g(AFI) represent factor models with 5 different levels each for Amount Requested and Amount Funded by Investor. The error term e represents all sources of unmeasured and unmodeled random variation in Interest Rate. Our final regression model appeared to remove most of the non-random patterns of variation in the residuals.

We observed a highly statistically significant ($P < 2e\text{-}16$) association between Interest Rate and FICO Range. A change of one unit of FICO Range corresponded to a change of b1 = -0.08807 on the Interest Rate (95% Confidence Interval: -0.09083, -0.08532).

Conclusions:

Our analysis suggests that there is a significant, negative association between Interest Rate and FICO Range. Our analysis estimates the relationship using a linear model relating FICO Range to Interest Rate. Even so, there appears to be a strong relationship between the two variables. We also observed that other variables such as Amount Requested and Amount Funded by Investor are associated with both loans Interest Rate and FICO Range. Including these variables in the regression model relating Interest Rate to FICO Range improves the model fit, but does not remove the significant negative relationship between the variables.

References:

1. Wikipedia "Interest Rate" page. URL: http://en.wikipedia.org/wiki/Interest_rate. Accessed 02/14/2013.

2. Wikipedia "FICO score" section page. URL: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States#FICO_score. Accessed 02/15/2013

3. R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 1/31/2013

4. Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 936. Wiley, 2012

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

6. The percentage of consumer's gross income that goes toward paying debts (http://en.wikipedia.org/wiki/Debt-to-income_ratio). Accessed 02/15/2013

7. The abbreviation for the U.S. state of residence of the loan applicant (http://www.50states.com/abbreviations.htm#.UQ8hxFp2PKo). Accessed 02/15/2013

8. The total amount outstanding all lines of credit (http://www.ehow.com/about_7550001_revolving-credit-balance.html). Accessed 02/15/2013

9. Inquiries in the Last 6 Months .This is the number of such authorized queries in the 6 months before the loan was issued (http://www.myfico.com/crediteducation/creditinquiries.aspx). Accessed 02/15/2013