1 Analysis Write-Up

1.1 Introduction

Lending Club is an online financial community that brings together borrowers and investors replacing the high cost and complexity of bank lending with a faster way to borrow and invest.

Borrowers can apply for a loan online and get an instant rate quote. Lending Club claims that the interest rate of these loans is determined on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores.

Here we performed an analysis to determine if there was a significant association between the interest rate of the loan and the characteristics of borrowers. Using exploratory analysis and standard multiple regression techniques we show that there is a significant relationship between interest rate and FICO score [1], even after adjusting for confounders such as the loan length and the amount funded.

Our analysis suggests that lower loan rate is associated with higher FICO score.

1.2 Methods

1.2.1 Data Collection

For our analysis we used a sample of 2,500 peer-to-peer loans issued through the Lending Club. The data were downloaded, using the R programming language [2], from here:

https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

Data were downloaded on November 7, 2013. No information were available on the time frame of the data.

The code book for the variables in the data set is available here:

https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf

1.2.2 Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the sample 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. determine the terms used in the regression model relating interest rate to FICO score

1.2.3 Statistical Modeling

To relate interest rate and FICO score we performed a standard multivariate linear regression model [3]. Model selection was performed on the basis of our exploratory analysis and knowledge of the relationship between interest rate and FICO score as found in Lending Club help pages [4].

1.2.4 Reproducibility

An RStudio [5] project is available at GitHub:

https://github.com/maurotrb/data-analysis-2013-project1

The repository contains the R markdown files [6] and the csv data file necessary to reproduce the analyses.

1.3 Analysis

1.3.1 Data quality

The loan data used in this analysis, composed of 2500 observations, contains information on the amount requested and funded, interest rate, loan length, loan purpose, debt to income ratio, state of the borrower, home ownership type, monthly income, FICO score (in ranges), open credit lines, revolving credit balance, inquiries in the last 6 months, and employment length.

We identified a few missing data in the monthly income, in the open credit lines, in the revolving credit balance, and in the inquiries in the last 6 months. Employment length has a significant number of missing data: 254 observations. Since those variables were not used for the final analyses, the quality of the data set was not affected by the missing values.

Six observations had the amount funded with values out of range: 0 and -0.10 dollars. We converted those values to missing data. During the exploration we found out that amount funded is a confounder for the linear regression model, so those 6 observations were excluded from the final analyses. The final data set was composed of 2494 observations.

1.3.2 Regression analysis

Interest rate distribution has a median of 13.10% (percentage points [7]) with a minimum of 5.42% and a maximum of 24.90%.

We explored the relation of interest rate with the other variables and found out that, apart from the obvious negative correlation with FICO score, there are significant positive correlation with loan length (expressed in months) and amount funded (expressed in US dollars).

We first fit a regression model relating interest rate to FICO score. To make the analysis easier, we transformed the FICO ranges present in the data set to numeric values, calculating the median value of each range (eg. 660-664 to 662).

The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including potential confounders such as loan length `LL` and amount funded `AF`. Amount funded was transformed in a factor variable dividing values into 3 ranges (quantiles). Exploratory analysis showed that 3 ranges were enough to be used in the model.

Our final regression model was:

```
IR = b0 + b1 * FS + f(LL) + g(AF) + e
```

where b0 is an intercept term and b1 represents the change in interest rate percentage points associated with a change of 1 unit of FICO score at the same loan length and amount funded range. The terms f(LL) and g(AF) represent factor models with 2 and 3 different levels each for loan length and amount funded range. 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 score. A change of 1 unit FICO score corresponded to a change of b1 = -0.08679 in the interest rate (95% Confidence Interval: -0.08922, -0.08436). So for example, for two loans with the same length and amount funded range, we would expect an interest rate for a borrower with a FICO score between 805 and 809 (807 median) to be 12.59 percentage points lower than one for a borrower with a FICO score between 660 and 664 (662 median) [8].

1.4 Conclusions

Our analysis suggests that there is a significant, negative association between interest rate and FICO score. Our analysis estimates the relationship using a linear model relating FICO score with percentage points of interest rate. We also observed that other variables such as loan length and amount funded are associated with both interest rate and FICO score. Including these variables in the regression model relating interest rate to FICO score improves the model fit, but does not remove the significant negative relationship between interest rate and FICO score.

Limitations in the analysis come from not knowing the time frame of the data set. Data could span months or years, during which loans criteria could have changed. While FICO score would remain probably the main predictor for interest rate, other confounders could result more significant with data classified by time periods.

1.5 References

The present analysis is based on the example analysis given in the Data Analysis course, from Coursera.

Other references:

1. FICO score. URL: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. Accessed 11/15/2013
2. The R Project for Statistical Computing. URL: http://www.R-project.org/.
3. Multiple and logistic regression. Chapter 8 of OpenIntro Statistics, 2nd edition. URL: http://www.openintro.org/download.php?file=os2_08&referrer=/stat/textbook.php. Accessed 11/17/2013
4. Interest Rates and How We Set Them. Lending Club help pages. URL: https://www.lendingclub.com/public/how-we-set-interest-rates.action. Accessed 11/17/2013
5. The R Project for Statistical Computing. URL: http://www.rstudio.com/ide/.
6. R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 11/15/2013
7. Percentage point. URL: http://en.wikipedia.org/wiki/Percentage_point. Accessed 11/17/2013
8. 12.59 percentage points computed as `807-662 * -0.08679`