Thomas M. Alisi
Jeff Leek
Data Analysis
18.2.13

INTEREST RATE IN LOANS IS ASSOCIATED TO FICO SCORE, AMOUNT REQUESTED AND LOAN LENGTH

Write-up

Introduction

A credit score in the United States is a number representing the creditworthiness of a person, the likelihood that person will pay his or her debts. [1]

The best-known and most widely used credit score model in the United States, the FICO score is calculated statistically, with information from a consumer's credit files. The score is sold by the FICO Company. [2] It provides a snapshot of risk that banks and other institutions use to help make lending decisions. Applicants with higher FICO scores might be offered better interest rates on mortgages or automobile loans as well as higher credit limit amounts.

Understanding the relationship between the FICO score and the Interest rate of a loan, in addition to other parameters describing the loan, can help us describing the nature of loans and, with further investigation, would lead to a prediction model for interest rate.

Using exploratory analysis and multiple regression techniques, we show that not only the FICO score affects the Interest rate of a loan, but also two other additional parameters, being the length of the loan and the amount requested. Our analysis clearly shows that increasing the two latter parameters (amount requested and lenght of loan) has the effect of increasing the Interest rate, regardless of higher FICO scores.

Methods

Data collection

For this analysis we used the loans data available from here: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda There is a code book for the variables in the data set available here: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf (both datasets and codebook where provided by Prof. Leek for this Assignment). The data has been downloaded on 11th of February 2013 consist of a sample of 2,500 peer-to-peer loans issued through the Lending Club [3].

Exploratory analysis

A first analysis of the data shows that there are only 7 NA values in the whole dataset. Analyzing details of NA values shows that they are related to 2 observations, so it is safe to ignore them since they do not represent a significant amount overall.

Further exploration of the dataset shows that several variables are expressed as factors, hence they have been converted to numerical values so as to provide a more flexible dataset for manipulation and analysis. The variables converted to numerical values are: Interest rate, FICO range

and other variables that were used during exploration but did not lead to significant correlation with the subject of this study.

Plotting a histogram of the Interest rate once converted to numeric value, shows that the distribution is lightly skewed with a higher percentage of smaller values. This distribution almost matches the histogram plotted for the Amount requested, showing an obvious relationship between lower amounts that implies lower interest rates.

Plotting values of the Amount funded against the Amount requested shows an almost straight line with a limited amount of lower values for the funded amounts. This makes our analysis almost replicable with values of the amount funded variable instead of the amount requested.

A detailed analysis and scatter plots of Interest rates against FICO range, shows that there are patterns associated to Loan lenghts and Amount requested. Other variables of the dataset does not seem to present clear patterns.

The exploratory analysis was thus used to (1) identify missing values, (2) verify quality and structure of data and (3) identify the variables used for the regression analysis related to interest rate and FICO score.

Statistical modeling

To relate Interest rate and FICO range in loans, we performed a multivariate regression model. The terms used for our final analysis are based on the results obtained during the exploratory analysis, and showing that there is clearly a pattern related to amount requested and loan length, while other variable do not show the same clear relation with the quantities analyzed.

Analysis

Our analysis started with a linear regression model [4] that takes into account the interest rate (IR) as outcome with the only FICO range as variable, the model can be defined as

$$IR = b0 + b1*f(FICO.Range)$$

We observed a significant dependency between the two variables ($P < 2e\text{-}16$) with an increase of one unit in the FICO range resulting in a decreased interest rate of 0.08. The residuals showed patterns of non-random variation, related to both the loan length and amount requested.

We attempted to explain those patterns by fitting models including potential confounders, and ended up with a model that takes into account the covariates mentioned in our exploratory analysis, hence the final model is:

$$IR = b0 + b1*f(FICO.Range) + b2*g(Interest.Rate) + b3*h(Loan.Length)$$

All these varaibles show significant dependency with the interest rate ($P < 2.2e\text{-}16$) and the analysis of residuals led to the removal of visual patterns related to the additional covariates, suggesting a better fit than our initial model and still describing the dependency of the Interest Rate on the terms used for this analysis.

Conclusions

We show that there is a significant association between interest rate and FICO score. We also show that variables affecting Interest rate are both the amount requested and loan length. Figure 1 panel (a) reports relationship between interest rate and amount requested, while panel (b) reports the analysis of residuals for the first regression model mentioned in the analysis section where a clear pattern emerges when plotting colours related to loan lengths. Our final regression model shown in panel (c) has an improved fit, with a more compact analysis of residuals with almost no patterns related to loan lenght. The residuals coloured by ranges of amount requested do not show any pattern hence they have not been reported in this analysis.

Potential problems of this regression model are: (1) other variables with weaker patterns may affect the relationship between FICO and interest rate, since the FICO score is a statistical value depending on some of the terms presented in this dataset and also (2) the results presented might be a little skewed due a misinterpretation of model coefficients and confidence intervals, mainly because of the lack of strong skills in data analysis, but we honestly hope that the outcome of this course will eliminate this covariate factor.

References

1. Wikipedia page, Credit score in the United States, accessed 17 Feb 2013 - http://en.wikipedia.org/wiki/Credit_score_in_the_United_States

2. Fair Isaac Corporation (FICO), official website, accessed 17 Feb 2013 - http://www.fico.com/en/Pages/default.aspx

3. Lending Club association, official websited, accessed 18 Feb 2013 - https://www.lendingclub.com/home.action

4. Wikipedia page, Linear regression, accessed 18 Feb 2013 - http://en.wikipedia.org/wiki/Linear_regression