Joseph Elliott

Data Analysis Assignment 1

Influences on Loan Interest Rate in Peer-to-Peer Lending

1 Introduction

The Lending Club is a peer-to-peer lending organization that provides loans for debt consolidation, paying off credit cards, home improvements, and other purposes (Lending Club, 2013). The interest rate at which the loan is issued is determined by many factors about the borrower's financial situation, including the borrower's monthly income, employment history, previous credit history, length of the loan term, and the amount of the loan. In particular, the borrower's FICO score, or credit score, has a strong influence on a borrower's perceived financial health. The FICO score is calculated using information about a borrower's credit history including their payment history, percent of available credit in use, the length of credit history, types of credit utilized, and frequency of inquiries about the credit file (Credit Score in the United States, 2013; Lending Club, 2013).

Determining interest rate of a loan is of interest to both the borrower and the lender. Loans are typically provided at lower interest rates to borrowers that are less likely to default on the loan. Thus a borrower's financial health is a good indicator of the interest rate they will receive. However, determining the interest rate is a complex process with many financial factors playing a role.

In this analysis, 2500 loans from The Lending Club will be analyzed to determine the relationship between interest rate and 13 other variables that indicate financial health. The FICO score is expected to play a primary role in determining the interest rate, but other variables may also exhibit a strong influence. Quantifying the effect these other variables have on the interest rate is the subject of this analysis.

A standard multivariate regression was performed that confirmed that higher FICO scores are associated with lower interest rates. In addition, exploratory analysis and regression determined that the length of the loan term, the amount requested, and the purpose of the loan also are related to the interest rate. This analysis will can be beneficial for borrowers or lenders in determining an appropriate interest rate for peer-to-peer loans.

2 Methods

2.1 Data Collection

The data set from The Lending Club was provided through the Coursera Course website.

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

And was downloaded on February 10, 2013 using the R programming language (The R Project for Statistical Computing, 2013). The data set consisted of 14 variables including:

- Amount Requested
- Amount Funded by Investors
- Interest Rate (the outcome)
- Loan length
- Loan purpose
- Debt to Income Ratio
- State
- Home Ownership
- Monthly Income
- FICO score range
- Number of open credit lines
- Revolving credit balance
- Number of credit inquiries in the last 6 months
- Employment length

2.2 Data Processing

As mentioned, the loan data provided is a sample of 2500 data points. To make the data set easier to work with, some cleaning processes were performed on the data set. Variables for interest rate and debt-to-income ratio were converted to numeric values between 0 and 1. For clarity values of interest rate in the text will use percentage notation. The FICO score ranges were converted to numeric values at the average of the range given. There were seven "NA" values; these observations were removed.

2.3 Statistical modeling

The statistical modeling carried out in this project consisted of standard multivariate regression using least-squares. This technique attempts to reduce the error and minimize the distance between the data points and the regression (Holman, 2001).

3 Results

Exploratory analysis was performed for the numeric variables to determine variables that could possibly influence the interest rate. The median interest rate was 13.11% with ranges from 0.542 5.42% to 24.89%, which is a normal range of values. Interest rate was bimodally distributed peaking around the median and also around 7%. The FICO score showed a right skew, with values between 642 and 832,

with a median at 682. It appears that there is a 'cutoff' for loan funding, as there are very few loans with FICO scores below 660. Performing a log transform on this data did not appear to even the distribution, so FICO score was not transformed. However, the Amount Requested was heavily right skewed, and a log10+1 transform was able to help normalize that distribution.

This analysis revealed a strong correlation between FICO score and interest rate. There was also a large difference in the median interest rate between 36 month loans (12.12%) and 60 month loans (16.49%). Other variables that appeared to be correlated with interest rate were debt-to-income ratio, amount requested, amount funded, and loan purpose.

A regression model was fit to four variables (FICO score, Amount Requested, Loan Purpose and Loan Length), to explain the variation in Interest Rate. The final correlation is:

$$Interest\ Rate = b_0 + b_1 1og_{10}(AmountRequested) + b_2 FICO + f(LoanLength) + g(LoanPurpose) + b_3 FICO * AmountRequested + b_4 AmountRequested * LoanLength$$

In this correlation, b0 is the intercept term, b2 is the change in interest rate with Log(Amount Requested), and b2 is the change in interest rate with FICO score. The term f(LoanLength) represents a factor with 2 levels (36 and 60 months), and the term g(LoanPurpose) represents a factor with 14 levels. There was a highly statistically significant correlations between FICO score and interest rate, as well as loan length and interest rate. Loan Purpose was significant ($p<0.05$) for two levels, "moving" and "other". There was a statistically significant correlation between Log(Amount Requested) and Interest Rate ($P=0.0014$). Two of the factors Two interaction terms between Log(Amount Requested):FICO and Log(Amount Requested):Loan Length were included which resulted in an improvement in the distribution of the residuals and an increase in the R^2 correlation coefficient to 0.7415. These two terms were also statistically significant.

The resulting correlation coefficient relating Interest Rate and Log(AmountRequested), b1=0.11 (95% Confidence Interval: 0.052, 0.163), where Amount Requested is in US dollars. Relating Interest Rate to FICO score, b2=-4.31e-4 (95% Confidence Interval: -7.44e-4, -1.17e-04). Figure 1 shows the relationship between the variables.

Joseph Elliott

Data Analysis Assignment 1

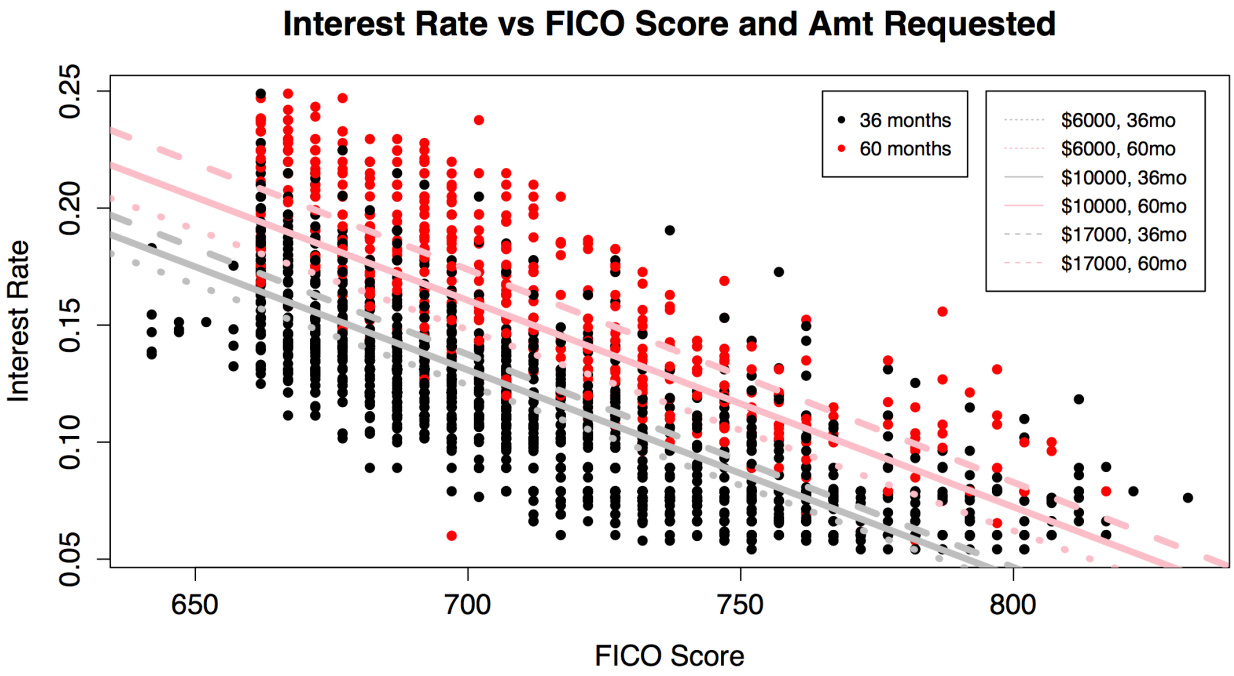

Figure 1. This figure is a scatter plot of the loan data plotting Interest Rate vs FICO score. Regression lines depicting interest rates combinations of Amount Requested (at the quartiles) for levels for Loan Length show how different interest rates can result for borrowers with a similar FICO score.

Several of these variables are considered confounders, in particular loan length, amount requested, and loan purpose. All are related, as larger loan amounts typically take require a longer time period to pay back, and the loan purpose can often have an influence on the amount requested. This was clearly seen in the exploratory analysis. Funding a wedding or vacation typically requires less money than an expense such as a car purchase or mortgage. Additionally, amount requested and amount funded appeared to be very tightly correlated.

In conclusion, the FICO score, Amount Requested, Loan Purpose, and Loan Length appear to correlated with Interest Rate. The association between Amount Requested and Loan Purpose is particularly strong, although these variables are confounders. This information may be used to help explain the variation in interest rates among borrowers. In particular, among borrowers with the same FICO score, borrowers requesting larger sums or longer loan terms may experience higher interest rates.

4 Works Cited

Credit Score in the United States. (2013 *йил* 5-February). Retrieved 2013 12-February from Wikipedia: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States

Holman, J. P. (2001). *Experimental Methods for Engineers (Vol. 7th).* Singapore, Singapore: McGraw-Hill.

Lending Club. (2013). Retrieved 2013 12-February from https://www.lendingclub.com/home.action

The R Project for Statistical Computing. (2013). Retrieved 2013 12-February from http://www.r-project.org/