**Title: Interested rate is associated with amount requested and loan length besides FICO score.**

**Introduction:**

Loans are necessary for conducting investments if the own capital is not sufficient. The loan is generally provided at a cost, referred to as interest on the debt, which provides an incentive for the lender to engage in the loan [1]. One variable that plays a key role in determining the interest rate for a given loan is the FICO score. The FICO score is calculated statistically, with information from a consumer's credit files [2]. In general, applicants with higher FICO scores might be offered better interest rates on loans or higher credit limit amounts. The FICO company has disclosed the following components which besides others calculate a person's credit score:

* Payment history
* Ratio of current revolving debt to the total available revolving credit or credit limit
* Length of credit history
* Types of credit used
* Recent searches for credit

Here we performed an analysis to determine if, besides the FICO score, there was a significant association between interest rate and amount requested and loan length. Using exploratory analysis and standard multiple regression techniques we show that there is a significant relationship between those variables. Our analysis suggests that a higher amount requested and a longer loan length result in a higher interest rate if two persons have the same FICO score.

**Methods:**

*Data Collection*

For our analysis we used the loans data provided by Professor Jeff Leek. The data were downloaded from the Amazon Simple Storage Service [3] on November 17, 2013 using the R programming language [4].

*Exploratory Analysis*

Exploratory analysis was performed by examining tables and plots of the observed data. 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 to FICO range, amount requested, and loan length.

*Statistical Modeling*

To relate interest rate to FICO score, amount requested, and loan length we performed a standard multivariate linear regression analysis [5]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between interest rate and FICO score. Coefficients were estimated with ordinary least squares.

**Results:**

The loan data used in this analysis contains information on the loan’s final interest rate (IR, measured in percent), the amount requested (AR) and the amount funded by investors (AF, both measured in dollars), the loan length (LL, 2 factor variables), the loan purpose (LP, 14 factor levels), the debt to income ratio (DIR, measured in percent), the state (St, 46 factor levels) in which the loan was issued, the applicant’s home ownership status (HO, 5 factor levels), the monthly income (MI, measured in dollars), the FICO score (FR, 38 factor levels) given in ranges, number of open credit lines (CL), revolving credit balance (RCB, measured in dollars), the number inquiries in the last six month (In), and the length of employment (EL, 12 factor levels).

The original data set contained 2500 observations. Two observations included explicit missing values (‘NA’). Moreover, an exploratory analysis showed that 77 observations included an implicit missing value for EL. These were labeled as ‘n/a’ in the respective column. In addition 6 observation were identified which had either a negative or zero value for AF. All of those observations were omitted from the data set after the exploratory analysis leaving 2415 observations for the subsequent regression analysis.

The exploratory analysis confirmed the already known relationship between the FICO score and the interest rate: The higher the FICO range the lower the interest rate. Besides that an additional linear relationship between the interest rate and the amount requested was discovered by grouping the data according to the 38 FICO ranges (see **Figure 1**: Top Left Panel).

The distribution of amount requested was right skewed. A log base 10 transformation had only a minor impact to the performance of the linear regression and was therefore discarded. Subsequent analysis focus on the amount requested variable.

We first fit a regression model relating interest rate to FICO range and amount requested. The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including a potential confounder, namely loan length (see **Figure 1**: Lower Left Panel).

Our final regression model was:

IR = b0 + b1AR + b2AR\*LL + f(FR) + g(LL) + e

where b0 is an intercept term and b1 represents the change in interest rate in percent associated with a change of 1 unit in dollars for borrowers within the same FICO range and a loan length of 36 months, whereas b2 does the same as b1 but for a loan length of 60 months. The term f(FR) represents a factor model with 38 different levels. The term g(LL) represents a factor model with 2 different levels. The error term e represents all sources of unmeasured and un-modeled random variation in interest rate. Our final regression model appeared to remove most of the non-random patterns of variation in the residuals (see **Figure 1**: Lower Right Panel).

.

We observed a highly statistically significant (P < 2e-16) association between interest rate and amount requested. An increase of amount requested by 1 dollar corresponds to an increase of the interest rate by b1 = 1.29e-04 percent (95% Confidence Interval: 1.15e-04, 1.42e-04) if the loan length is 36 months. Another highly statically significant (P = 1.99e-04) association between interest rate and the interaction term ‘amount requested x loan length’ was observed. An increase of amount requested by 1 dollar corresponds to an increase of the interest rate by b2 = 4.55e-05 percent (95% Confidence Interval: 2.16e-05 6.96e-05) if the loan length is 60 months.

So for example, for two loans with a difference of 10.000 dollars being issued with a loan length of 36 months for borrowers within the same FICO range, we would expect the interest rate to be 1.29 percent higher for the bigger amount.

**Conclusions:**

Our analysis suggests that there is a significant, positive association between interest rate and amount requested (besides FICO range). Our analysis estimates the relationship using a linear model relating interest rate to amount requested and FICO range. We also observed that loan length is associated with both interest rate and amount requested. Including this variable in the regression model improves the model fit.

The exploratory data analysis only revealed one covariate candidate which in the end was highly statistically associated with the interest rate. The reason for that might be that most of the data set’s other variables are part of the original FICO score calculation (see Introduction) and therefore implied already.

**References**

1. Wikipedia “Loan” Page. URL <http://en.wikipedia.org/wiki/Loan>. Accessed 11/16/2013.

2. Wikipedia “Earthquake” Page. URL: <http://en.wikipedia.org/wiki/Credit_score_in_the_United_States>. Accessed 11/16/2013.

3. Loan data location. URL: <https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv>. Accessed 11/17/2013.

4. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: [http://www.R-project.org](http://www.r-project.org)

5. Wikipedia “Linear Regression” Page. URL <http://en.wikipedia.org/wiki/Linear_regression>. Accessed 11/16/2013.