Title: Lending Club's interest rate is predominantly associated with only 3 variables

Introduction:

Lending Club is a US peer to peer lending company. Lending club enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request [1].

To borrow from lending club, you must have a FIFO score of at least 660 and a debt-to-income ratio (excluding mortgage) below 35%. In addition, your credit history must show that you are a responsible borrower with 2 or more revolving accounts, with more than 36 months of credit history and with less than 7 credit inquiries in the last 6 months. Currently loan applications from 6 states are not accepted [2].

The purpose of this analysis is to identify and quantify associations between the interest rate of the loan and the other variables in the data set. In particular it is considered which variables have an important association with interest rate after taking into account the applicant's FICO score.

For the borrower it will be very interesting to learn that only 2 factors in addition to the FICO score do explain more than 78% of his interest rate, although he has to reveal to Lending Club more than ten other factors during his loan application.

Methods:

Data Collection

For the analysis a dataset of 2,500 loans data on https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda were downloaded on 2/10/2013 using the R programming language [3].
The same process was applied for the code book for the variables on the same date on https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf.

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 to interest rate.

Statistical Modeling

To relate the interest rate to all the other variables that are requested by the Lending Club a standard multivariate linear regression model [4] was developed. Model selection was performed on the basis of the exploratory analysis.
The adjusted R-square statistic measures the model's overall predictive power and the extent to which the variables in the model explain the variation in the dependent variable, the interest rate.
A forward selection was chosen which improved step by step by adding the variable with highest contribution to the adjusted R-squared statistic until the improvement steps were less than 1%. Finally interactions between the factors in this minimal adequate model and polynomial terms were used to improve the model further [5].
Regressive Diagnostics verified the model assumptions.

Reproducibility

All analyses performed in this manuscript are reproduced in a R markdown file. 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.rda might change based on the date.
Due to security concerns the R code wont be attached to this manuscript [6].

Results:

Loans data used in this analysis contain information about the interest rate determined by Lending Club associated with information about the amount requested, funded, the length of the loan and its purpose, debt to income ratio, state, home ownership, monthly income, FICO range, open credit lines, revolving credit balance, inquiries in the last 6 months and employment lenth.
A logarithmic transformation was required for the monthly income due to its right skewed distribution.

The exploratory analysis indicated already the major correlations between interest rate, FICO score, amount requested and the length of the loan. Figure 1 shows graphically the scatterplot of these relations which also correspond to the final regression model described further below.

We first fit a regression model relating interest rate to Fico score after transforming the FICO range to a numeric value by taking the mean of the range. This transformation allows the borrower easily to work directly with his FICO score in the final regression model and not to search for the relevant ranges of this data set and it allows the usage of polynomial regression. Figure 2 shows the residuals versus fitted for the linear model with Fico score as a predictor.

A forward selection was chosen which improved step by step the adjusted R-squared statistic until the improvement was less than 1%. Adjusted R-squared jumped from 0.50 with FICO as

the unique predictor to 0.66 when Amount requested as second predictor was added. With loan length as the third predictor the linear model reached its peak with 0.74. Adding any of the other variables of the data as a predictor did not show any improvement above 0.1.,
Squared polynomial regression applied on FICO showed another considerable jump to 0.78. Figure 3 demonstrates the improvement of residuals versus fitted for the final polynomial model when compared to figure 2, FICO as the only predictor.
Cubical polynomials as well as interactions did not improve the result.

The final regression model for the interest rate IR was:

$$IR = b_0 + b_1(FICO) + b_2(Amount\ Requested) + b_3(Loan\ Length) + b_4(FICO\text{^}2) + e$$

where b_0 is an intercept term and $b_1 + b_4$ represent the interest change of one FICO unit for the same requested amount and the same loan length. b_2 and b_3 are the coefficents for the predictors Amount Requested and Loan Length. The error term e represents all source of unmeasured and unmodeled random variation in the interest rate.

A highly statistically significant (P = 2.2e-16) association between IR and the three predictors can be observed.

A positive change of one FICO unit corresponds to a negative change of $b_1 + b_4 = 0.86\%$ in interest rate assuming that the other predictors are constant. Similarly an increase of the requested amount of 10000 USD would result in an increase of the interest rate of 1.38% and a two year loan prolongation would result in an increase of the interest rate of 3.30%.

Confounding can be excluded, since all the chosen three predictors are not dependent from another variable outside the model which is a pre-requisite for a lurking variable, since a confounder influences by definition the dependent as well as the independent variables, the predictors. FICO is setup by another organization and loan length and amount requested are the decision of the borrower.
There were no missing values in the data for the three predictors used in the final model. Regression diagnostic plots confirmed the validity of the model assumptions.

Conclusions:

My analysis suggests that there is a significant, negative association between interest rate and FICO score and a positive association between interest rate and amount requested and loan length. These three variables contribute to predict the variation of the interest rate up to 78%.

It became evident that the borrower should maximize always his FICO score, minimize his loan amount or break it up into several loans and minimize the loan length.

After this study the borrower might not understand why he has to provide so many data to Lending Club whereas only a small amount determine his interest rate.

The reason might lie in the fact that the investors need those data to calculate their risk before granting the loan.
Thus it would be of great interest for investors to perform a similar study on Lending Club data to optimize their investments. For this purpose additional data on default rates should be taken into account as well [7].

References

1. Wikipedia “Earthquake” Page. URL: http://en.wikipedia.org/wiki/Lending_Club Accessed 2/15/2013.
2. Lending Club “What are the basic requirements from borrowers” Page. URL: http://www.lendingclub.com/kb/index.php?View=entry&EntryID=186 Accessed 2/15/2013.
3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org Accessed 2/15/2013.
4. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 936. Wiley, 2012.
5. Darlene R. Goldstein, EPFL, http://lausanne.isb-sib.ch/~darlene/as/ModelSelection.pdf Accessed 2/15/2013.
6. Coursera, Jeff Leek, https://class.coursera.org/dataanalysis-001/human_grading/view/courses/294/assessments/4/submissions Accessed 2/15/2013.
7. Lending Club, What is the default rate on Lending Club loans?, URL: http://www.lendingclub.com/kb/index.php?View=entry&EntryID=81 Accessed 2/15/2013.