Title: High Loan Interest Rates are commonly associated with lower FICO ranges

Introduction:

Loans can provide us with immediate access to money that we may need for various purposes. They are generally provided at a cost, referred to as interest on the debt, which provides an incentive for the lenders to engage in the loans [1].

The most important factor to assess the credit worthiness of a person in the United States is by means of a scoring system developed by FICO, previously known as Fair Isaac Corporation. It uses a risk-based system to determine the possibility that the borrower may default on financial obligations to the mortgage lender. [2] Besides the FICO score, other factors including debt-to-income ratio, amount requested, and home ownership can also contribute to the disbursement of loans.

Understanding the relationship of interest rates to the various factors can help us better understand how lenders make their decisions. Here, I performed an analysis to determine if there was a significant association between interest rates and FICO ranges. Using exploratory analysis and standard multiple regression techniques, I show that there is a significant relationship between interest rates and FICO ranges, even after adjusting for important confounders such as Amount Funded by Investors, Debt to Income Ratio, and Home Ownership. My analysis suggests that increased interest rate is associated with lower FICO range. My results suggest that there are relatively few interest rates that are not influenced by FICO scores.

Methods:

Data Collection

For my analysis, I used loan-related data originally available at the company Lending Club's website [6]. Lending Club is an online financial community that brings together creditworthy borrowers and savvy investors so that both can benefit financially [6]. Lending Club data containing 2500 samples were downloaded from Coursera's Data Analysis Assignment 1 webpage [7] on February 9, 2013 using the R programming language [3].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. I 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 interest rate and FICO range.

Statistical Modeling

To relate interest rates to FICO ranges, I performed a standard multivariate linear regression model [4]. Model selection was performed on the basis of exploratory analysis and observing the P-value and F-value for each covariate. Coefficients were estimated with ordinary least squares and standard errors were calculated using standard asymptotic approximations [5].

Analysis and Results:

The loan data used in this analysis contains information that measured the Amount Requested (AR, in US dollars), Amount Funded By Investors (AF, in US dollars), Interest.Rate (IR, measured in percentage), Loan.Length (LL, in months), Loan.Purpose (LP), Debt To Income Ratio (DIR, measured in percentage), State (S), Home Ownership (HO), Monthly Income (MI, in US dollars), FICO Range (FICO), Open Credit Lines (CL), Revolving Credit Balance (CB, in US dollars), Inquiries in the Last 6 Months (INQ), and Employment Length (EL, in total years) [7]. I identified 7 missing (NA) values in the data set I collected and all measured variables were observed to be inside the standard ranges. I noticed 3 distinct outlier values in the Monthly Income data.

I also identified the following confounders that had a relationship with both the interest rate and FICO range: Amount Requested, Amount Funded by Investors, Debt To Income Ratio, Home Ownership, Open Credit Lines, and Inquiries in the last 6 months.

Most loans had interest rates in the 10 to 15% range. There were also a number of interest rates in the 6 to 8% range. The distribution of FICO ranges was mostly left skewed, with several values being in the ranges between 675 and 700.

I first fit a regression model relating interest rate to FICO range. The residuals showed patterns of non-random variation with residual standard error 2.946 on 2498 degrees of freedom. I attempted to explain those patterns by fitting models including potential confounders. I then observed the P-value and F-values by relating each covariate in the table below to interest range. I left out the variables that were statistically not significant ($P > 0.001$).

Covariate	P-Value	F-Value	Selected (Yes/No)
Amount.Requested (AR)	<2e-16	310.19	**YES**
Amount.Funded.By.Investors (AF)	<2e-16	320.87	**YES**
Loan.Length (LL)	<2.2e-16	546.55	**YES**
Loan.Purpose (LP)	0.0019	9.6654	NO (P-value> 0.001)
Debt.To.Income.Ratio (DIR)	<2e-16	77.259	**YES**
State (S)	0.679	0.1713	NO (P-value> 0.001)
Home.Ownership (HO)	0.000174	14.136	**YES**
Monthly.Income (MI)	0.5396	0.3764	NO (P-value > 0.001)
FICO.Range (FICO)	<2e-16	2527.5	**YES**
Open.CREDIT.Lines (CL)	6.17e-06	20.523	**YES**
Revolving.CREDIT.Balance (CB)	0.00225	9.3557	NO (P-value > 0.001)
Inquiries.in.the.Last.6.Months (INQ)	<2e-16	69.548	**YES**
Employment.Length (EL)	0.5303	0.394	NO (P-value > 0.001)

Table 1 – Table indicating the covariates that were chosen for the final regression model

My final regression model was:

$$IR = b_0 + b_1\,FICO + f(AR) + g(AF) + h(LL) + i(DIR) + j(HO) + k(CL) + l(INQ) + e$$

where b_0 is an intercept term and b_1 represents the change in interest rate (IR) in percentage associated with a change of 1 unit for FICO range (FICO) assuming the same values for Amount Requested, Amount Funded, Loan Length, Debt to Income Ratio, Home Ownership, Open Credit Lines, and Inquiries in the Last 6 Months.

The terms f(AR), g(AF), h(LL), i(DIR), j(HO), k(CL), and l(INQ) represent regression models for Amount Requested, Amount Funded, Loan Length, Debt to Income Ratio, Home Ownership, Open Credit Lines, and Inquiries in the Last 6 Months. The error term e represents all sources of unmeasured and unmodeled random variation/uncertainty in interest rate. My final regression model appeared to remove several non-random patterns of variation in the residuals.

I observed a highly statistically significant (P = 2.2e-16) association between interest rate and FICO range with a lower residual standard error of 2.047 on 2489 degrees of freedom. A change of one unit in FICO range corresponded to a change of b_1 = -0.43 percentage in interest rate (95% Confidence Interval: -0.44, -0.42).

Conclusions:

My analysis suggests that there is a significant, positive association between interest rate and FICO range. My analysis estimates the relationship using a linear model relating interest rate to FICO range. There appears to be a strong relationship between the two variables. I also observed that other variables such as Amount Requested, Amount Funded, and Loan Length are associated with both interest rate and FICO range. Including these variables in the regression model relating interest rate to FICO range improves the model fit, but does not remove the significant positive relationship between the variables.

While this analysis is an interesting first step, it is based on a limited sample of loans from the data available for the assignment. A larger collection of representative loans may be more appropriate for understanding the relationship between interest rates and FICO ranges. My analysis may be of interest to scientists and students seeking to better understand loans. But for money lending institutions in the real world, the analysis may involve other complex variables that have not been considered in this assignment.

References:

1. Wikipedia “Loan” Page. URL: http://en.wikipedia.org/wiki/Loan. Accessed 2/9/2013.
2. Wikipedia “Credit Score” Page. URL: http://en.wikipedia.org/wiki/Credit_score. Accessed 2/16/2013
3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org, Accessed 2/10/2013.
4. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 936. Wiley, 2012.
5. Ferguson, Thomas S. *A Course in Large Sample Theory: Texts in Statistical Science*. Vol. 38. Chapman & Hall/CRC, 1996.
6. Lending Club. URL: http://www.lendingclub.com/, Accessed 2/10/2013.
7. Lending Club data. URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv**.** Accessed 2/9/2013.