An Analysis of Factors Influencing Interest Rate in a Body of Peer to Peer Loan Data

Introduction

Interest is a fee charged for borrowing assets [1]. Among money lenders, the interest rate is a percentage of the borrowed amount charged over time. This rate has widespread ramifications, affecting not only persons attempting to finance personal purchases or small businesses, but also impacting the nation's supply of money and banks' ability to lend money [2]. The interest rate for a personal loan is adjusted to reflect the volatility of the loan; that is, the risk that the loan will not be repaid. As such, it is in the interest of borrowers to understand what factors will impact their ability to get a loan, and in the interest of lenders to understand how their competitors are evaluating risk and attracting potential customers. In this analysis, we will examine data regarding loans from the Lending Club (https://www.lendingclub.com/home.action) to identify any factors that correlate with increasing interest rate.

Methods

The data analyzed describe 2500 peer-to-peer loans issued through the Lending Club. These loans are a subset of those publicly available from the Lending Club's website. The data used for the analysis are mirrored on a public data share on Amazon.com [3]. The data are formatted for use with the R statistical programming language, and all analyses were performed using R [4].

Exploratory analyses were performed with the intent to validate the data, identify required data transformations for analysis, and identify important variables for further regression analyses. Techniques used in the exploratory analysis included creating univarite and multivariate plots, tables, and matrices.

The relationship between the interest rate and the other factors included in the data set were modeled using multivariate linear regression, via the least-squares method. The final model was selected based on prior knowledge of information used to set interest rates, as well as the results from the exploratory analysis [5].

Results

The data contained information on the amount requested by the borrower; the amount funded by investors; the interest rate; the length of the loan; the purpose of the loan; the borrower's debt to income ratio; the borrower's state of residence; the borrower's home ownership status; the borrower's monthly income; the borrower's FICO score (reported as a range); the number of credit lines the borrower has open; the current balance across all of the borrower's credit lines; the number of credit inquiries from the last six months on the borrower's history; and the length of time the borrower has been employed [6].

The average interest rate across all loans was approximately 13%. The purpose of the majority of the loans was debt consolidation (52%), with the next largest reason being credit card payoff (17%). 78% of the loans were for a duration of 36 months, while the remainder (12%) were for 60 months. The FICO scores of the borrowers were skewed right with a sharp drop below 660-664, with a median of

700-704, 25th percentile of 680-684, and a 75th percentile of 725-729. A histogram of the applicants' FICO scores are visible in Figure 1a.

We fitted a linear regression model to the data respective to the applicant's Interest rate (INTR), FICO score (FICO), loan amount (LAMO), loan length (LLEN), monthly income (MINC), revolving credit balance (RCRE), number of inquiries in the past six months (INQU). The final model was:

$$INTR = b_0 + b_1 \cdot FICO + b_2 \cdot LAMO + b_3 \cdot LLEN + b_4 \cdot MINC + b_5 \cdot RCRE + b_6 \cdot INQU$$

where b_0 is an intercept term and the other b_i terms are coefficients. This model significantly improved upon the initial linear regression model relating only FICO score and interest rate, by removing much of the non-random variation.

Data about the regression are located in Table 1. The FICO score was the variable the most strongly correlated with the interest rate (-8.469e-04, P-value < 2e-16). For every incremental increase in FICO score, the interest rate would lower by approximately 0.089%. This corresponds to a difference across consecutive FICO groups of approximately 0.432%.

Conclusion

This model shows that there is a strong correlation between a loan's interest rate, and the applicant's FICO score, loan amount, loan length, monthly income, revolving credit balance, and number of credit inquiries in the previous 6 months. Upon close examination of the model, especially the Residuals vs Fitted Values graph shown in Figure 1b, while the variance of the graph is mostly accounted for, there is still a pattern that is unaccounted for. This may suggest that there is an external variable that may be confounding this analysis, or that thee is a polynomial term that has not been included in this analysis. Looking at the Residuals versus Leverage graph in figure 1c indicates that there are some outliers that are swaying the model's parameters more than would be expected. These data points appear to be valid, and more research into evaluating the properties of similar data points would be warranted.

We have provided insight into how the Lending Club evaluates risk in its loan applicants by quantifying the amount of risk the Lending Club attributes to each term in our regression model. This information could be used by other banks to audit their lending policies, or by loan applicants who are interested in managing how risky they appear to banks. This model does have some issues with fit, and further research is needed into how this model would stand up to data from other banks.

Reference

1. Interest. *Wikipedia: The Free Encyclopedia.* http://en.wikipedia.org/wiki/Interest. Accessed 2013-11-17.
2. Interest Rate. *Wikipedia: The Free Encyclopedia.* http://en.wikipedia.org/wiki/Interest_Rate. Accessed 2013-11-17.
3. Data used in this analysis. https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda. Accessed 2013-11-17.
4. R Core Team (2012). “R: A language and environment for statistical computing” http://www.R-project.org.
5. Hamburg, Morris. “10. Multiple Regression and Correlation Analysis.” In: *Statistical Analysis for Decision Making*. 3rd Ed. New York: New York 1983.
6. *Lending Club Statistics*. https://www.lendingclub.com/info/download-data.action. Accessed 2013-11-17.

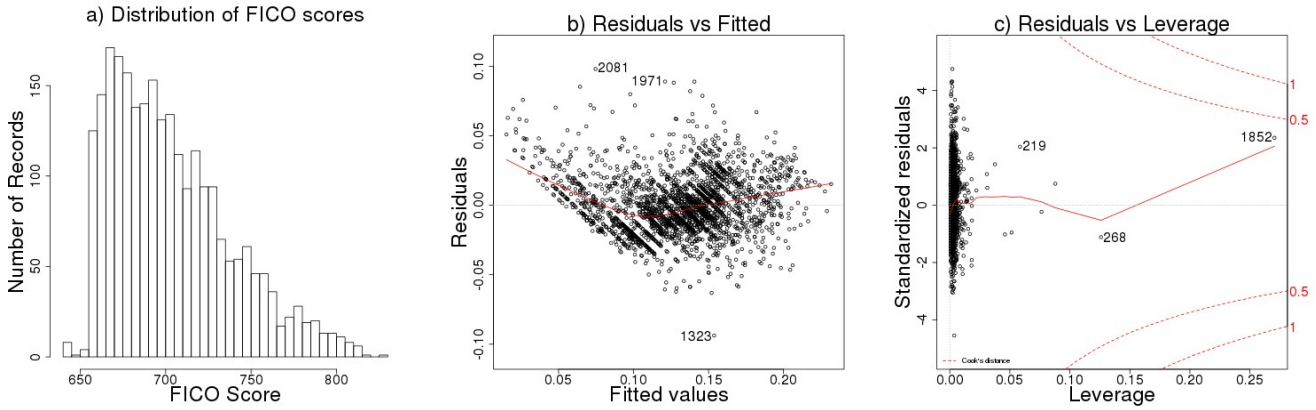

Figure 1: a) A histogram displaying the right-skewedness of the FICO score distribution. b) A graph of residuals versus fitted values that clearly displays the presence of systematic error in the data. c) A graph of standardized residuals versus leverage that clearly displays the extent of outlier influence over the regression model.

Residuals				
Min	1Q	Median	3Q	Max
-0.093763	-0.013847	-0.001589	0.012257	0.098041
Coefficients				
	Estimate	Std. Error	T-value	Pr(>|t|)
Intercept (b0)	60668e-01	8.620e-03	77.361	< 2e-16
FICO score (b1)	-8.649e-04	1.198e-05	-72.204	< 2e-16
Loan Amount (b2)	1.515e-06	6.460e-08	23.459	< 2e-16
Loan Length (b3)	1.334e-03	4.601e-05	28.994	< 2e-16
Monthly Income (b4)	-2.669e-07	1.194e-07	-2.236	0.02547
Revolving Credit Balance (b5)	-6.730e-08	2.464e-08	-2.731	0.00636
Inquiries in the last 6 months (b6)	3.390e-03	3.381e-04	10.028	< 2e-16
Multiple R-squared: 0.7563			Adjusted R-squared: 0.7557	
F-statistic: 1289 on 6 and 2491 DF			P-value	< 2.2e-16

Table 1: Summary of Multiple Regression