Title: Peer-to-peer loan interest rates are affected by more than just FICO credit ratings

Introduction:

The Lending Club is an online financial community which facilitates peer-to-peer loans. [1] Each loan request is assigned a 'loan grade' based on a number of different factors, and then assigned a particular interest rate [2]. One important variable for determining the interest rate is the credit-worthiness of a person, which is generally determined by the FICO credit score, and is itself calculated from a customer's credit files [3]. A higher FICO score generally represents lower risk for banks and lending institutions and often results in an individual getting better (lower) interest rates [3].

In this analysis, we performed an analysis of a sample of 2,500 loans made by the Lending Club to identify and quantify the relationship between the interest rates offered with the FICO score, and at the same time, to assess whether any other variables played an important role in determining the interest rate. Using exploratory analysis and standard multiple regression techniques, we determined that while the FICO score has a very significant relationship to the interest rate, three other factors – the length of time the loan is for, the amount funded by investors and the number of open credit lines – were all significantly correlated to the interest rate. This suggests that individuals who share the same FICO credit rating might very well be offered different interest rates by the Lending Club based on these other factors.

Methods:

Data Collection

For our analysis we used the data on 2500 loans made by the Lending Club. The data was downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda on February, 16 2013 using the R programming language [4].

Exploratory Analysis

We conducted exploratory analysis by examining summaries of the loans data with plots and tables. This was done in order to identify transformations to make on the raw data, and used to remove a few fields with missing data and transform character/range data into factors or numbers to simplify analysis. Following this, each of the variables in the original data were plotted against the interest rate, using scatterplots and boxplots [5].

Statistical Modeling

In order to determine how important each of the remaining variables were in explaining the interest rate, we performed a standard multivariate linear regression model, with coefficients were estimated with ordinary least squares and standard errors were calculated using standard asymptotic approximations [5]. The variables included in the regression model were based on the exploratory analysis described above.

Results:

The loans data we analysed contained data on 14 variables for each loan made. These were the interest rate, the FICO rating range, purpose of loan, length of the loan, amount of loan requested, amount of loan funded by investors, monthly income, the Debt-to-Income Ratio, number of open credit lines, amount of revolving credit, state, housing ownership status, employment length and no. of past credit inquiries that had been made about the person in the past 6 months.

Initial exploratory analysis indicated that 5 of the original variables (i.e. Purpose of Loan, State, Employment Length, Housing Status and No. of Past Credit Inquiries Made) had no or extremely low correlation with interest rate and were then removed from further analysis. In addition, two individuals with missing values in their income/credit history were removed from the data set. The FICO range for each individual was transformed by assigning a different number to each range to produce a new FICO rating for the purposes of this analysis, with lower FICO ranges assigned a lower rating.

The regression model was therefore initially calculated based on 8 variables selected by exploratory analysis shown above. Based on this calculation, it was found that the coefficients for three of the variables, the Debt to Income Ratio of an individual, Monthly Income and was not statistically significant in explaining interest rate at a p-value of 1%. Therefore these variable were also removed from the analysis, and the regression analysis was performed again with 5 variables. It was also suspected that the amount requested and the amount funded might act as confounders and this was indeed found to be the case – since the correlation between the two variables was greater then 95%. Therefore, the variable ‘Amount Requested’ was also removed, which changed the significance of the effect of Amount Funded on interest rate by a great margin. The final regression model can be expressed as shown in Table 1.

Table 1. Multiple Regression Analysis of Variables Impacting Interest Rate

Interest Rate = Intercept + β1*FICOnum+ β2*AmountF + β2*OpenC + β3*length + error

Residuals:

Min	1Q	Median	3Q	Max
-9.9717	-1.4101	-0.1643	1.2558	10.3261

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.765e+01	1.490e-01	118.443	< 2e-16	***
FICOnum	-4.397e-01	6.069e-03	-72.453	< 2e-16	***
AmountF	1.436e-04	6.083e-06	23.603	< 2e-16	***
OpenC	-3.555e-02	9.566e-03	-3.717	0.000206	***
length	3.294e+00	1.114e-01	29.556	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.104 on 2493 degrees of freedom
Multiple R-squared: 0.7468, Adjusted R-squared: 0.7463
F-statistic: 1838 on 4 and 2493 DF, p-value: < 2.2e-16

As seen in table 1, the FICOnum was the FICO Rating, AmountF was the amount of loan funded, OpenC was the number of open credit lines the individual possessed, and length was the loan length in months.

From table 1, it can be seen that the loan length had the highest effect on interest rate, followed by the FICO rating. The number of open credit lines had the next biggest effect, followed by the actual amount funded by investors for the loan. This suggests strongly that two individuals with the same FICO rating could get different interest rates for their loans, if for example, one individual requested a loan for 24 months and the other requested for 36 months, or if the amount requested/funded is different. Furthermore, the fact that the number of open credit lines made a significant impact on the interest rate shows that the Lending Club did not rely solely on the FICO ratings to determine the credit-worthiness of an individual but also relied on how many credit lines an individual had open to decide the

interest rate: in this case, an individual was likely to get a lower interest rate if they had more credit lines open.

Conclusion

Our analysis suggests that there is a significant positive association between the length of time and the amount of loan funded with the interest rate, and a significant negative association with FICO ratings and the number of open credit lines with the interest rate. The conclusions are based on a limited data sample of 2,500 loans funded, and may differ with a larger data set. It is also not known what was the time period during which these loans were funded, and it is possible that other factors – such as bank lending rates of the time, stock exchange values, etc. which are not captured in this data set might strongly influence the interest rate offered as well.

References

1. The Lending Club website. https://www.lendingclub.com/public/about-us.action. Accessed on 16-02-2013 at 10.33 pm (GMT+8)
2. The Lending Club website 'Rates and Fees'. https://www.lendingclub.com/public/rates-and-fees.action. Accessed on 16-02-2013 at 10.35 pm (GMT+8)
3. Wikipedia 'Credit score in the Uniter States'. http://en.wikipedia.org/wiki/Credit_score_in_the_United_States. Accessed on 16-02-2013 at 10.39 pm (GMT+8)
4. R Core Team (2012). "R: A language and environment for statistical computing." http://www.R-project.org.
5. Seber, George AF, and Alan J. Lee. *Linear regression analysis*. Vol. 936. Wiley, 2012.