Increased interest rates in peer-to-peer lending are associated with increased loan length and requested loan amount.

Introduction:

Peer-to-peer lending is the practice of lending money to individuals, typically in the form of unsecured personal loans, outside the normal banking infrastructure [1]. The most important variable in determining an interest rate for personal loans in the United States is the creditworthiness of the borrower [2]. Two other variables (beyond creditworthiness) that also have a significant association with determining the interest rate of a peer-to-peer loan are the amount and term (length) requested by the borrower.

Understanding the relationship of interest rate to the size and length of a loan can help a borrower determine whether to take multiple small loans or a single larger loan. In a peer-to-peer lending environment, such as Lending Club [3], where the origination fee is in many cases smaller for shorter loans [4], applying for multiple smaller loans could provide the borrower with overall lower interest payments.

Using exploratory analysis and standard multiple regression techniques we show that there is a significant relationship between interest rate and the loan amount requested and loan length, even after adjusting for the borrower's creditworthiness. Our analysis suggests that both an increased amount requested and longer loan lengths are associated with an increased interest rate.

Methods:

Data Collection

For our analysis we used data on 2,500 peer-to-peer loans issued through the Lending Club [3]. The data were provided by Professor Jeff Leek as part of the Data Analysis [5] course offered by Coursera.org on January 22, 2013. The data were downloaded February 4, 2013 using the R programming language [6].

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) identify unusual values, (3) generally verify the quality of the data, and (4) to determine which variables have an important association with interest rate after taking into account an applicant's creditworthiness.

Statistical Modeling

To test for an association between interest rate and the amount requested and loan length we used a multivariate linear regression model [7]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between FICO Score [2] - which is a measure of creditworthiness - and other variables such as debt to income ratio, revolving credit balance and recent credit inquiries.

Reproducibility

All analyses performed in this manuscript are reproduced in the R files exploratory.R and final.R (not submitted as part of this assignment).

Analysis/Results:

The loans data used in this analysis contains information on the amount requested by the borrower (Amount.Requested), the amount funded by investors (Amount.Funded.By.Investors), the interest rate of the loan (Interest.Rate), the length of the loan (Loan.Length), the purpose of the loan (Loan.Purpose), and the borrowers debt to income ratio (Debt.To.Income.Ratio), FICO score (FICO.Range), monthly income (Monthly.Income), state of residence (State), number of open credit lines (Open.CEDIT.Lines), credit inquiries (Inquiries.in.the.Last.6.Months), employment length (Employment.Length) and revolving credit balance (Revolving.CREDIT.balance).

We removed 2 observations that included missing values that could not be reliably reconstructed, leaving us with 2,498 of the original 2,500 observations. Our data also included some unusual values for the amount funded by investors (e.g., -1 and 0). Since the analysis does not make use of amount funded variable, and since the other variables associated with the effected observations appear reasonable, we chose not to eliminate those observations from the data.

Since a borrower's creditworthiness is already known to be a significant factor in determining interest rates we performed a baseline analysis using a linear regression model that fit interest rates using the FICO score of the borrower.

$$\text{Interest.Rate}_1 = b_0 + b1*(\text{FICO.Range}) + e,$$

Where b0 in an intercept and b1 represents the change in interest rate associated with a change of 5 points in the borrowers FICO score. The error term e represents all sources of unmeasured variation in interest rates not accounted for in our model. Our base regression model suggests that FICO score has a significant impact on interest rate (p-value: < 0.0001) and that the FICO score variable alone explains more than half of the total variation in the interest rate in the data set (Multiple R-squared: 0.5451, Adjusted R-squared: 0.5383).

The positive correlation of interest rate with a borrower's FICO score is not surprising since the FICO score is the primary metric by which lending institutions measure the risk associated with lending money.

A second baseline analysis was created using a multiple linear regression model that fit interest rates using all variables except the amount funded by investors. This more complete regression model suggests that beyond the FICO score the variables that have an impact on interest rate (p-value: < 0.1) are: the amount requested by the borrower, loan length, monthly income, number of open credit lines, and the number of recent credit inquiries. This more complete model explains around 80% of the total variation in the interest rate in the data set (Multiple R-squared: 0.8056, Adjusted R-squared: 0.7995). However, there is a known association between FICO score [2] and number of open credit lines and the number of recent credit inquiries; meaning this model is almost certainly more complex than necessary.

The more complete multiple linear regression model aside, the only variables beyond the FICO score that showed a strong correlation with the interest rate were the amount requested by the borrower and the length loan (see figure 1 – left and center panels).

We fit a more limited (simpler than the more complete model) regression model relating interest rate to: FICO score, amount requested and loan length:

$$\text{Interest.Rate}_1 = c_0 + c_1*(\text{FICO.Range}) + c_2*(\text{Amount.Requested}) + c_3*(\text{Loan.Length}) + e,$$

Where c_0 in an intercept and c_1, c_2 and c_3 represents the change in interest rate associated with a change of 5 points in the borrowers FICO score, a change of $5,000 in the amount requested and a change between a 36 and 60 month loan length. As in the previous models, the error term e represents all sources of unmeasured variation in interest rates not accounted for in our model.

Our limited model accounted for approximately 79% of the variation in the interest rate in the data set (Multiple R-squared: 0.79, Adjusted R-squared: 0.7862), and each variable in the model showed a highly statistically significant associations with the interest rate (p-value: < 0.0001 for all variables). This regression model suggests that: 1) the Lending Club bases their interest rates at least partly on the amount of the loan request and on the length of time a borrower wishes to take to pay back the loan, and 2) that the other variables provided in the data set are provide no discernible information beyond FICO score, amount requested and loan length. In other words, our more limited model is just as good at predicting interest rates as the more complete model presented earlier.

The residuals from our limited regression model (FICO score, amount requested and loan length) appear to be fairly normal (see figure 1 – right panel). The curvature and deviation from the normal distribution of the residuals at the boundaries –especially in upper quintiles - suggest that our data has heavier tails than normally distributed data [9]. Our exploratory analysis showed that the amount requested distribution is positively skewed (right tail is longer), but no reasonable data transformation was found to fix this anomaly. This could be an underlying cause for the deviation from the normal distribution in the Q-Q Plot in figure 1 – right panel.

It should be noted that some lower FICO scores appear to have less statistical significance. The FICO score data was binned and some bins representing FICO scores below 700 showed little if any statistical significance (p-value > 0.05). This is an anomaly in our data that may suggest that for lower FICO scores there are other variables significantly contributing to the interest rate. We have not explored any possible variables (measured or unmeasured) that account for this anomaly, but it would be worth investigating in a more in depth data analysis.

Conclusions:

Our analysis suggest that beyond the expected association with between interest rates and FICO score there is also a significant association between the interest rate assigned to a loan and: 1) the amount requested by the borrower and 2) the amount of time the borrower wishes to take to repay the loan. Our analysis also suggests that many of the measured variables in the data Lending Club data set are also conflated into the FICO score; explaining why we see no difference in explained variation between the more complete and limited models.

While the length of a loan is clearly associated with the interest, it is a well known practice in institutional lending to offer lower interest rates for shorter term loans. In fact the Lending Club website shows that they have a specific policy where longer loans in general imply higher interest rates.

While the association between FICO score and loan length are not surprising, that our analysis shows a positive association between the amount requested in a loan and the interest rate is interesting. Nowhere on the Lending Club website is it mentioned that the amount of the loan has any bearing on the eventual interest rate. This may be explained by the nature of the peer-to-peer lending, where an individual investor may try to lower their risk by only accepting a higher interest rate for a larger capital exposure. Finally, while this analysis only shows an associative and not causal relationship, it does suggest that a possible strategy for borrowers is to apply for multiple smaller loans when the origination fee for an individual loan is not prohibitive. In the case of Lending Club the origination fee is a fixed percentage of the loan, meaning that the borrower would suffer no penalty by following this strategy).

References

1. Wikipedia “Peer-to-peer lending” Page. URL: http//en.wikipedia.org/wiki/Peer-to-peer_lending (Accessed: 2/12/2013)
2. Wikipedia “Credit score in the United States” Page. URL: http//en.wikipedia.org/wiki/Credit_score_in_the_United_States (Accessed: 2/13/2013)
3. Lending Club. URL: http//lendingclub.com/home.action
4. Lending Club – Rates and Fees. URL: https://www.lendingclub.com/public/rates-and-fees.action
5. Data Analysis class Assignment 1 Loan Data. URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.rda. (Accessed 2/4/2013)
6. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org
7. Multiple Linear Regression. URL: http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm
8. Goodness-of-Fit Statistics. URL: http://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html
9. Online Statistics Education: An Interactive Multimedia Course of Study. URL: http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html