Title: Loan Length and Loan Amount Affect Lending Club Interest Rates for Applicants with Same Credit Scores

Introduction:

Peer-to-peer lending involves the practice of money lending between individuals (usually strangers), without involvement of traditional financial institutions, such as banks [1]. It is usually conducted online via a lending company such as Lending Club, Corp. which acts as a broker between parties. In this type of lending system, borrowers have the opportunity to pay less interest than they would on typical credit cards or bank loans, while lenders could get higher returns than they would in more traditional types of investments. There are, however, many types of risks associated with this practice. Risks to the investor, for example, involve the type of loan and the borrower's creditworthiness score or FICO score (lower scores reflect poor credit history). Higher risk is reflected in higher interest rates, which are undesirable for borrowers as they increase the cost of the loan. The purpose of this analysis is to identify and quantify associations between the interest rate of the loan and the various characteristics of the loan and their applicants, after adjusting for the effect of the applicant's FICO score. These associations are important, as they help give insight as to how peer-to-peer lenders determine a loan's interest rate.

Methods:

Data collection

Data on peer-to-peer loans was obtained from the Lending Club [3], downloaded as a csv file on 11/06/2013 [4], and analyzed using R Statistical Computing Software [5].

Exploratory Analysis

Exploratory analysis was performed by examining tables, plots and correlations of the observed data to determine the variables that appear to influence most the Interest Rate.

The dataset included 2,500 observations (or samples) of the loan's Interest Rate and a total of 13 predictor variables: Amount.Requested, Amount.Funded.by.Investors, FICO Range, Loan.Length, Loan.Purpose, State, Monthly.Income, Open.CREDIT.Lines, Inquiries.in.the.Last.6.Months, Home.Ownership, Debt.to.Income.Ratio, Revolving.CREDIT.Balance, and Employment.Length.

We identified 7 missing values (NA's) in the original dataset. All 7 values were replaced by the mean of their corresponding column. To reduce the complexity in the modeling, the FICO.Range variable, originally recorded as a range "Min-Max", was converted from categorical variable to numeric variable by using the Min value for each observation. Similarly, the "%" character in the variable "Interest Rate" and the variable "Debt.To.Income.Ratio" was dropped and the observations converted to "numeric".

Statistical Modeling

Multivariate linear regression analysis [6] was performed to identify and quantify associations between the “Interest Rate” (response or dependent variable) and the various predictor variables (independent variables) in the dataset, and determine the most significant predictors of the response Interest Rate, after adjusting for the effect of the applicant’s FICO score (confounder). Multivariate linear regression demonstrates the unique contribution of each categorical or continuous independent variable to the variability in the Interest Rate.

First, a simple linear regression model was used to estimate the Interest rate (IR) as a function of the FICO score (FICO) and establish a baseline. Second, using the “Forward Selection Method” [7], selected predictor variables were added to the model sequentially and their respective contribution quantified using multiple linear regression. The results were interpreted using the following statistical significance tests:

a) the Adjusted-R^2 statistic, which is a measure of how much the response’s variability is due to its relationship with the independent variable(s). In general, the value of the Adjusted-R^2 statistic will increase as significant variables are added to the model and decrease if unnecessary terms are added.
b) the p-value, which, if less than 0.05, gives strong evidence that a variable’s independent effect on the response variable Interest Rate is considered statistically significant.

Finally, we confirmed our findings using the “Backward Elimination Method”, which involved including all predictors in the model and eliminating the variables with large p-values sequentially.

Results:

1. <u>Establish Baseline: Effect of FICO Score on Interest Rate</u>

The effect of the predictor variable FICO Score on the response variable ”Interest rate” (IR) was analyzed using a simple linear regression model :

$$IR = \beta_1 FICO + \varepsilon$$

The resulting fit shows that the coefficient $\beta_1 = -0.0846$, indicating that a one unit increase in the applicant’s FICO score corresponds to a 0.0846 unit reduction (in %) in the loan’s Interest rate. The p-value for the variable FICO is <0.0001 indicating that this variable is highly significant, and the adjusted R^2 value for the model is found to be 0.5026. This result indicates that the FICO variable alone can predict up to 50% of the variability in the Interest Rate.

Figure 1 displays the Interest Rate versus FICO Score, color grouped by the seven levels of Interest Rates as defined by the Lending Club. The graph shows that higher FICO scores correspond with lower Interest Rates, while lower FICO scores correspond with higher Interest Rates.

Next, the other variable's effect on the response variable Interest Rate (IR) were tested by adding them to the regression model, one at a time. A multivariate linear regression analysis was performed to determine the most significant predictors of the response variable Interest Rate, after adjusting for the effect of the applicant's FICO score (confounder). A variable's effect on the response variable IR is considered statistically significant if a) the new variable coefficient's p-value is less than 0.05, and b) the adjusted R^2 value increases by at least 0.1 with respect to the baseline adjusted R^2 value of 0.5026. This increment is how much the variability is uniquely explained by the independent variable, while controlling for the other variables.

2. Effect of the Loan's Length on Interest Rate with adjustment for FICO score.

The Loan.Length variable has two possible values: 36 months and 60 months.

A multivariate regression model including Loan.Length as a variable along with the FICO variable increased the adjusted R^2 value to 0.6896 (an increase of 0.187 from the baseline model's adjusted R^2 value). When including the second level interaction term Loan.Length*FICO, the adjusted R^2 value increased to 0.6928 (an increase of 0.1902 from the baseline model's adjusted R^2 value).

The β_2 coefficient for the variable Loan.Length is positive indicating that 60-months loans have the effect of increasing interest rates as opposed to 36-month loans for applicants with the same FICO score. The corresponding p-value is less than 0.001 indicating that the predictor Loan.Length is statistically significant in predicting the response Interest Rate. The model suggests that 70% of the variability in Interest Rate can be attributed to the two variables FICO Score and Loan.Length.

Figure 2 displays the Interest Rate versus FICO Score, color grouped by the 2 levels of loan length (Red for 60 months and Black for 36 months). The graph shows that those who apply for a longer loan (60 months in Red) have higher Interest Rate than those who apply for a shorter loan (36 months in Black).

3. Effect of the Amount Requested on Interest Rate with adjustment for the FICO score.

A multivariate regression model including Amount.Requested as a variable along with the FICO Score variable increased the adjusted R^2 value to 0.6564 (an increase of 0.154 from the baseline model's adjusted R^2 value). When including the second level interaction term Amount.Requested*FICO, the adjusted R^2 value increased to 0.661.

The β_2 coefficient for the variable Amount Requested is positive indicating that an increase in the Amount Requested corresponds to an increase in the loan's Interest rate. The corresponding p-value is less than 0.001 indicating that the variable Amount Requested is statistically significant in predicting the Interest Rate. This result shows that both variables FICO Score and Amount Requested help explain ~ 66% of the variability in Interest Rate.

Figure 3 displays the Interest Rate versus FICO Score, with five groupings of amount requested (Light Blue for the highest Amount Requested and Black for the lowest Amount Requested). The graph shows that those who requested higher amounts (light blue dots) will have higher Interest Rates than those who requested lower amounts (Black dots).

It is interesting to note that the variable Amount Funded by the Investors shows similar significance on the response Interest Rate. This is not surprising since the Amount Requested is highly correlated with the Amount Funded by the Investors (Pearson's correlation coefficient=0.9698).

A multivariate regression model including the three variables Loan Length, Amount.Requested and FICO Score and a third level interaction term (FICO*LoanLength*AmountRequested) increased the adjusted R^2 value to 0.7482 (an increase of 0.2456 from the baseline model's adjusted R^2 value), indicating that ~75% of the variability in Interest Rate can be attributed to these 3 variables: FICO Score, Loan Length and Amount Requested.

a. Other effects and testing

When including all 13 predictor variables in a multivariate regression model, the adjusted R^2 value increased to 0.7621 (an increase of 0.2595 from the baseline model's adjusted R^2 value R^2 value). This suggests that all 13 predictors can explain ~76% of the variability in the response Interest Rate, but that there may be additional factors and interactions that contribute to the Interest Rate that are not measured and modeled. After removing the statistically insignificant terms following a "Backward Selection Method", we performed a multivariate regression on the remaining 5 covariates: FICO Score, Loan Length, Amount Requested, the variable Open.Credit.Lines and the variable Inquiries.in.the.Last.6.Months. The resulting adjusted R^2 value decreased to 0.7565, which is less than 1% different from the Adjusted R^2 value obtained with the 3 main variables: FICO Score, Loan Length and Amount Requested. This result confirms that the three most important predictor variables of the Interest Rate are FICO Score, Loan Length and Amount Requested. Alone these variables predict more than 75% of the Interest Rate.

Conclusions

Using multivariate linear regression analysis, we found that the Length of the loan and the Amount of the loan (Amount Requested or Amount Funded) were the most significant predictors of Interest Rate, while controlling for the applicant's FICO score. Loans with a length of 60 months were shown to increase Interest Rates compared to loans with a length of 36 months. Applicants who requested higher loan amounts received higher Interest Rates than those who requested lower loan amounts. Other variables like Open.Credit.Lines and Inquiries.in.the.Last.6.Months were also significant predictors, but with lesser effects.

Multivariate linear regression analysis is a good first order approach to determining the most significant variables that contribute to a response. It assumes, however, the relationships between the variables to be linear in first order. While this may be true in some cases, many types of relationships between variables may be non-linear and/or require more complex modeling. The main advantage of multiple regression models, however, is to be able to include control variables, a way to measure the impact of any given variable while adjusting for the impact of another variable.

References

[1] Wikipedia "Peer-to-peer lending" Page. URL: http://en.wikipedia.org/wiki/Peer-to-peer_lending. Accessed 11/06/2013.

[2] MyFICO.com. URL: http://www.myfico.com/CreditEducation/articles/. Accessed 11/06/2013.

[3] The Lending Club. URL: https://www.lendingclub.com/. Accessed 11/06/2013.

[4] Project Data Set. https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. Accessed 11/06/2013.

[5] The R Project for Statistical Computing. URL: http://www.r-project.org/. Accessed 11/06/2013.

[6] "Chapter 7: Modeling Relationship of Multiple Variables with Linear Regression". URL: http://www.pearsonhighered.com/assets/hip/us/hip_us_pearsonhighered/samplechapter/0205863728.pdf. Accessed 11/06/2013.

[7] StepWise regression. URL: http://en.wikipedia.org/wiki/Stepwise_regression. Accessed 11/06/2013.