Primary Variables Associated with Lending Club Loan Interest Rates

1.0 – Introduction

Lending Club is a business that facilitates peer lending [1]. Investors provide capital to Lending Club, which then lends that capital to borrowers at varying interest rates. Each investor's capital is distributed across a number of loans, although investors are able to specify their risk tolerance, preferred loan types, and other parameters [2].

The question of interest is:

What are the principal factors that determine the interest rate set by Lending Club for a particular borrower?

In particular, it is desirable to distinguish between two borrowers who have similar credit ratings.

Understanding what factors contribute to Lending Club interest rate is helpful for both borrowers and lenders. For borrowers considering applying for a Lending Club loan, knowledge of what variables influence the interest rate they are likely to receive can save them money if they are able to modify those factors to obtain a lower interest rate. Prospective lenders can use the same information to evaluate Lending Club's Loan Grades [3], making it easier to assess whether a given grade of loans fall within each lender's individual risk tolerance.

2.0 – Methods

2.1 Data Collection

The dataset for this analysis contained 2,500 loans issued by Lending Club, each with 14 variables such as the interest rate that the borrower received, their FICO score [4], income, and credit balance.

The dataset was downloaded in CSV format from a public Amazon S3 host [5] on February 10, 2013 using Firefox 18.0.2. The following variables were identified:

- The loan amount requested by the borrower ($0 - $35,000)
- The loan amount issued by Lending Club ($0 - $35,000)
- The interest rate of the loan (approximately 5-20%)
- The term of the loan (36 or 60 months)
- The purpose of the loan (e.g. car, medical, debt consolidation, vacation)
- The borrower's monthly income
- The borrower's outstanding credit balance (including credit cards, lines of credit, etc) [6]
- The borrower's debt to income ratio [7]
- The borrower's home ownership status: whether they own their home, rent, or have a mortgage
- The borrower's state of residence (Lending Club only issues loans in the United States)
- The range of the borrower's FICO score, within 5 (e.g. a FICO score of 734 falls in the 730-734 range)

- The number of open lines of credit held by the borrower
- The number of loan inquiries made by the borrower in the last 6 months
- The duration of the borrower's current employment (not applicable if the borrower is unemployed or retired)

2.2 Exploratory Analysis and Transformations

All analysis was performed in the R programming language using RStudio 0.97.248. A simple data parser was written in R to clean up and transform the data. In parallel, tables and plots of the dataset were examined to identify appropriate transformations, problematic observations, and other factors.

- Monthly income was transformed with a base 10 logarithm to normalize its distribution, which was right skewed. The transformation was of the form *log(value + 1)*. Log transformation of other monetary values was tested, but did not enhance distribution normality or model performance.
- The purpose of the loan, the borrower's state of residence, and the duration of the borrower's employment were converted to factor variables.
- Because loan length was either 36 or 60 months, it was also converted to a factor variable.
- After importing it as a factor and comparing the means and 25th/75th percentiles of the interest rates associated with the borrower's employment length, all 77 of the loans with "not applicable" employment length entries were removed from the dataset. These entries did not exhibit significantly different qualitative behavior, and removing them simplified the exploratory analysis process by allowing employment length to be studied both as a factor and as a numeric value within a single cleaned/transformed dataset.
- Two values of debt to income ratio were undefined. These loans were removed from the dataset.
- The range of the borrower's FICO score was replaced with the numeric value of the low end of the score range. Because the score bins are ranged 0-4 and 5-9, this did not result in any borrowers being moved across a key FICO score threshold. The key FICO score thresholds are 600, 660, and 720 [8].

The resulting dataset contained 2,421 loans with 14 variables for each. Further exploratory analysis was then performed by plotting variables from the cleaned and transformed dataset to confirm distribution normality and identify possible correlations.

2.3 Inferential Analysis

Standard multiple linear regression [9] was applied iteratively to determine which variables other than FICO score were the largest contributors to explaining variance in interest rates.

3.0 – Results

A standard multiple linear regression was performed to identify the largest contributors to explaining variance in interest rates. FICO score was included in this regression by default, since it was known to be a factor and showed a clear negative correlation to interest rate (R-squared = 0.50), as expected. A two-factor model was constructed for FICO with each other variable, and the R-squared values were compared to determine which two-factor model explained the most variance.

This analysis identified the loan length as the second-largest contributor to explaining variance after FICO score, with a two-factor model R-squared of 0.690. The effect of loan length is shown in Figure 1-1: boxplots of interest rate by loan length for borrowers seeking 36-month or 60-month loans show a clear difference between their means and 25th/75th percentiles. The overlaid violin plots show the density distributions for the two groups of loans; the 60-month loans have a broader Interest Rate distribution centered on a higher rate.

Once loan length was identified as the second largest contributor to explaining variance, a second multiple linear regression was performed to identify other key contributors to interest rate. This resulted in a set of three-factor models, each including FICO score, loan length, and one other variable. The R-squared values of each model were then compared to determine which variable improved the model the most.

The amount requested was found to contribute the most to improving the model, with an R-squared = 0.745. Repeating this exercise with a fourth variable did not significantly enhance the model. The maximum R-squared with a fourth variable was 0.754, achieved by adding the number of inquiries in the last 6 months. This was not considered to be a sufficient improvement to merit including a fourth term in the model, which would further complicate model interpretation. Therefore, the final regression model was:

$$\mathrm{IR} = b_0 + b_1\,\mathrm{FICO} + b_2\,I_{60}\,\mathrm{LL} + b_3\,\mathrm{AR} + e$$

The model parameters are defined as:

- IR is the interest rate the borrower received
- FICO is the borrower’s FICO score
- LL is the length of the loan (36 or 60 months)
- AR is the amount requested, the monetary value of the loan requested by the borrower
- b_0 is an intercept term
- b_1 represents the change in Interest Rate associated with FICO score
- b_2 represents the change in Interest Rate if Loan Length is 60 months
- I_{60} is a conditional/dummy variable that is 0 if LL = 36 and 1 if LL = 60
- b_3 represents the change in Interest Rate associated with Amount Requested
- *e* is an error term representing sources of unmeasured or unmodeled variation.

A known and highly statistically significant inverse correlation was observed between the borrower’s FICO score and Interest Rate ($P < 2e-16$). An increase of 1.0 in FICO score resulted in a decrease of 0.0875% in Interest Rate (95% Confidence Interval: -0.0899, -0.0851). An increase of 60 in FICO score – enough to move a borrower across one of the key thresholds – resulted in an Interest Rate decrease of 5.25%.

A highly statistically significant correlation was observed between Loan Length and Interest Rate (P < 2e-16). Stepping Loan Length from 36 to 60 months resulted in an increase of 3.30% in Interest Rate (95% Confidence Interval: 3.08, 3.52).

A highly statistically significant correlation was also observed between the log of Amount Requested and Interest Rate (P < 2e-16). An increase of $1 in the Amount Requested resulted in an increase of 1.38e-4 percentage points in Interest Rate (95% Confidence Interval: 1.26e-4, 1.50e-4). For two borrowers with the same FICO score and Loan Length, a borrower requesting a $30,000 loan would be expected to receive an interest rate 1.38% higher than a borrower requesting a $20,000 loan.

A plot of the model residuals (see Figure 1-2) shows an error bias as Interest Rate approaches 20%. No version of the three- or four-variable model reduced or eliminated this bias significantly. This suggests that there are additional variables influencing Interest Rate that are not included in this dataset.

Larger loan datasets with additional variables included are available from Lending Club [10], including funded loans, loans available for funding, and declined loans. The funded loans data was downloaded on February 16th, 2013, using Firefox 18.0.2. This dataset includes 112,043 funded loans between 2007 and 2010 with 41 variables. Possible additional variables that may be correlated with Interest Rate include:

- Dates associated with the loan (application date, issue date)
- Number of loan delinquencies by the borrower in the past two years
- The borrower’s total number of credit lines (including ones that are now closed)
- Percentage utilization of open credit lines
- The monthly payment associated with the loan

In particular, loan issue date is often a factor in Interest Rate, due to variation in interest rates available from the United States Federal Reserve and primary lenders such as banks or mortgage lenders. After excluding 3,301 loans where either interest rate or issue date was missing, a scatterplot of loan issue date over time (Figure 1-3) shows that the maximum interest rates on Lending Club loans increased steadily between 2007 and 2013. This exploratory plot suggests that incorporating loan issue date into the model developed for this analysis may be helpful in reducing the model’s bias in prediction error for high interest rates.

4.0 – Conclusions

The analysis shows that for two borrowers with the same FICO score, there are significant positive associations between the term of the loan, the value of the loan requested by the borrower, and the interest rate the borrower ultimately received from Lending Club. A significant negative association between FICO score and interest rate received was also observed, as expected.

Lending Club reports some of these factors on their website under “Interest Rates” [3]. They use a proprietary model to assign a risk grade to a borrower using criteria that include (but are not limited to) the applicant’s credit score. Lending Club then applies modifiers to these risk grades based on the loan term and requested loan amount. These risk modifiers are shown on Tables 3 and 4 on the Lending Club Interest Rate page, and broadly align with the model in this analysis: a loan length of 60 months carries larger risk modifiers than amount requested in most cases.

The model relationships are all linear, whereas the risk modifiers Lending Club reports are discrete. The model could be revised to use Lending Club's risk modifiers directly, which would likely improve its accuracy.

Finally, it is clear that there are other unmeasured variables that influence the interest rate received by borrowers. Broadening this analysis to include the larger loan datasets available from Lending Club would allow these variables to be investigated. A preliminary review suggests that loan issue date may be a significant factor.

5.0 – References

[1] Lending Club main page. Accessed 2/15/2013.
URL: http://www.lendingclub.com/

[2] How Lending Club Works. Accessed 2/15/2013.
URL: https://www.lendingclub.com/public/how-peer-lending-works.action

[3] Lending Club: Interest Rates and How We Set Them. Accessed 2/15/2013.
URL: https://www.lendingclub.com/public/how-we-set-interest-rates.action

[4] Wikipedia "Credit Score in the United States" page, "FICO Score" section. Accessed 2/15/2013.
URL: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States#FICO_score

[5] Amazon S3 site for Data Analysis. Accessed 2/10/2013.
URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

[6] eHow: Revolving Credit Balance. Accessed 2/15/2013.
URL: http://www.ehow.com/about_7550001_revolving-credit-balance.html

[7] Wikipedia "Debt to Income Ratio" page. Accessed 2/15/2013.
URL: http://en.wikipedia.org/wiki/Debt-to-income_ratio

[8] Lending Club "Check Your Rate" page. Accessed 2/15/2013.
URL: https://www.lendingclub.com/borrowerc/applyForALoan.action

[9] Yale "Statistics 101" course page, 1997. Accessed 2/15/2013.
http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[10] Lending Club "Statistics" page. Accessed. 2/16/2013.
URL: https://www.lendingclub.com/info/download-data.action