Title: Which data drive the Interest Rates on the peer-to-peer loans issued through The Lending Club?

Introduction:

It is expected that loan applicants' creditworthiness is closely tied to their FICO score. Lending institutions use the FICO score as an indication of borrowers' likelihood to repay their debt. It is expected that the interest rate charged on loans are closely associated to borrowers' FICO score [1]. However, FICO scores are not the only data point that drives the interest rates on the peer-to-peer loans issued through The Lending Club. The interest rates on the 2,500 record data sample are associated with not only FICO scores, but also with the length of the loan and the purpose of the loan, as well as home ownership status and the number of credit inquiries in the last 6 months.

Understanding the relationship of the various borrower data collected by The Lending Club clarifies the factors involved with establishing interest rates for loans. An analysis was performed to determine if there was a significant association between interest rates and data other than the FICO score. Using exploratory analysis and standard regression techniques, relationships between a few other data points where identified. The results suggest that four data account for a portion of the interest rates that are charged.

Data points contributing to borrowers' assessed interest rate are [2]:
Loan.Length - The length of time (in months) of the loan (either 36 or 60 months).
Loan.Purpose - The purpose of the loan as stated by the applicant.
Home.Ownership - A variable indicating whether the applicant owns, rents, or has a mortgage on their home.
Inquiries.in.the.Last.6.Months - When a person applies for credit, they authorize the lender to "inquire" about their creditworthiness. This is the number of such authorized queries in the 6 months before the loan was issued (http://www.myfico.com/crediteducation/creditinquiries.aspx).

Methods:

Data Collection

This analysis used data consisting of a sample of 2,500 peer-to-peer loans issued through The Lending Club (https://www.lendingclub.com/home.action). The data were downloaded from https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv on Saturday, Feb 09, 2013 using the R programming language [3].

Exploratory Analysis

Exploratory analysis was performed by examining and quantifying observed data. Various data were transformed to facility measurement by scale and rank ordering. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms used in the regression model relating interest rates to FICO scores and the other data contributing to variations in interest rates.

Statistical Modeling

To relate interest rates to FICO and other data, standard linear regression models were performed, as well as analysis of variance (ANOVA) and Tukey multiple comparisons of means [4,5,6,7,8]. A correlation matrix of all variables was run and examined for expected and unexpected significant relations [9].

Results:

The Lending Club's loans data was apportioned to five data sets that met the following criteria:

- The first data set contained the entire sample set for comparison to the following subsets
- The next two subsets contained records that are consistent with the expected relationship between FICO scores and interest rates:
  - Subset 1: Records with Low FICO scores AND High Interest rates (see Figure 1)
  - Subset 2: Records with High FICO scores AND Low Interest rates
- The final two subsets contained records that are inconsistent (counter-intuitive) with the expected relationship between FICO scores and interest rates:
  - Subset 3: Records with Low FICO scores AND Low Interest rates
  - Subset 4: Records with High FICO scores AND High Interest rates (see Figure 1)

The raw data used in this analysis contains information on the interest rates formatted as percentages. A data point was created to quantify the whole number portion of the interest rate (for example, if the interest rate was 15.31%, the created value was 15 (stored in variable IntRatePrim). A similar approach was utilized to transform FICO ranges. A data point was created to quantify the beginning portion of the FICO range (for example, if the FICO range was 645-649, the created value was 645 (stored in variable FICOBeg). The data contained in the Inquiries.in.the.Last.6.Months variable was ranked into four levels "A","B","C","Z" relating to the best values (zero inquiries assigned to "A", one inquiry assigned to "B", two inquires assigned to "C", and all records with over two inquiries assigned to "Z") and stored in a created variable (InquiresRank). Both the original data points and the created data points were used in the analysis to measure the relationship between interest rates and other key contributors.

The distribution of income data was heavily right skewed. To improve the performance of linear regression techniques containing these data, a log base 10 transformation of monthly income was performed.

A regression model relating FICO scores to interest rates was fitted, as a starting point, to confirm the relationship. Subsequent regression models relating interest rate to loan length, loan purpose, home ownership and inquiries in the last 6 months were performed on the entire data set as well as the four subsets of data.

An association with a high statistical significance (P = <2e-16) was found in all the data sets (the entire data set as well as the four subsets of data) between interest rate and inquiries in the last 6 months ranked as "A", between interest rate and loan term, between interest rate and home ownership, and between interest rate and loan purpose (especially debt consolidation). A Tukey multiple comparisons

of means test was utilized to confirm the relationship between interest rate and loan purpose in which 92 difference loan purposes were evaluated with P-value most significant for records containing loan purposes of debt_consolidation-car, home_improvement-credit_card, major_purchase-credit_card, home_improvement-debt_consolidation, major_purchase-debt_consolidation and other-major_purchase.

Conclusions:

The analysis suggests that there is a significant association between interest rates on the 2,500 record data sample with not only FICO scores, but also the length of the loan, the purpose of the loan, home ownership and the number of credit inquiries in the last 6 months. The analysis estimated the relationship using a linear model to measure the association strength.

While this analysis is an interesting first step, in future analysis, attention should be given to the potential confounders in the data set, perhaps including monthly income and other factors that may influence the FICO score.

References

1. Wikipedia Page "Credit score in the United States" URL: http://en.wikipedia.org/wiki/Credit_score_in_the_United_States Accessed 2/17/2013
2. The Lending Club loans codebook Page Accessed 2/9/2013 URL: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf
3. R Core Team (2012). "R: A language and environment for statistical computing." URL: http://www.R-project.org
4. Description of Statistical Formulas in Excel Page URL: http://www.comfsm.fm/~dleeling/statistics/excel.html Accessed 2/16/2013
5. More on Exploring Correlations in R Page URL: http://www.r-bloggers.com/more-on-exploring-correlations-in-r/ Accessed 2/13/2013
6. The Elements of Statistical Learning (2nd edition), Hastie, Tibshirani and Friedman (2008). Springer-Verlag. 763 pages, online version URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/ Accessed 2/13/2013
7. Tidy data, Hadley Wickham, September 23, 2011 URL: http://vita.had.co.nz/papers/tidy-data.pdf Accessed 2/13/2013
8. Advanced Data Analysis from an Elementary Point of View, Cosma Rohilla Shalizi, Spring 2013, Last LATEX'd Wednesday 30th January, 2013
9. How To Analyze Simple Two-Way and Multi-Way Table, Correspondence Analysis Page URL: http://www.statsoft.com/textbook/correspondence-analysis/ Accessed 2/13/2013
10. Getting Started with Lattice Graphics, Deepayan Sarkar, http://lattice.r-forge.r-project.org/Vignettes/src/lattice-intro/lattice-intro.pdf Accessed 2/16/2013
11. Basic Introduction to ggplot2 Page URL: http://www.r-bloggers.com/basic-introduction-to-ggplot2/ Accessed 2/13/2013

NOTE: Fully analysis and documentation available via GitHub upon request.