Prepared by: Mary Anne Plano
January 14, 2014

## Introduction

The Lending Club is an online loan provider which offers both lenders and borrowers the opportunity to either borrow or loan money [1]. A number of factors are usually considered when determining the interest rate of a loan; often these are encapsulated in the borrower's credit score.[2] In this analysis we investigate what other factors affect the interest rate assigned to a particular loan issued through the Lending Club. The potential for confounds are quite high with the provided data, as the information collected can also be utilized to determine an applicant's credit score. So, a preliminary model without confounders is investigated and then the interdependence of factors is investigated. The results indicate that the most important factors in determining the interest rate of a loan are the FICO score, the loan length, the state of residence and the amount funded by investors. With a simple linear model based solely on single factors 76.5% of the variation in interest rate can be explained. A more complex model containing interactive terms is only able to account for an additional 1% of the variation. Further modeling efforts could be focused on capturing the interrelationship of the variables provided and utilize more sophisticated non-linear models.

## Methods

### *Data Collection*

For this analysis we used a sample of 14 variables for 2500 peer-to-peer loans issued through the Lending Club[1]. The data were downloaded from the provided website [3] using the R programming language[4]. All analyses were performed using the R programming language and a script illustrating the commands utilized is posted online [5].

### *Exploratory Analysis*

An initial survey of the data revealed the need to transform data provided as characters and factors into numeric data. Also, the credit score data was reported in a range of width 4, so the low score of the range was selected for use and analysis. An exploratory analysis was performed by examining tables and plots of the provided data. Results of this analysis were used to verify the quality of the data, identify missing values and determine the terms to use in the regression model relating the interest rate charged to the various applicant parameters.

### *Statistical Modeling*

To relate interest rate to loan properties a standard multivariate linear regression model was applied [6,7]. Model selection was performed on the basis of the exploratory analysis and by creating a maximal model without interactions and reducing to the simplest model based on an

analysis of variance. After that the addition of interaction terms were investigated on the basis of mutual dependency of the most important variables.[8]

Analysis

The loans data used in this analysis contains information on 14 variables: the amount requested (AR), amount funded by investors (AF), interest rate charged (IR), length of the loan (LL), loan purpose (LP), debt to income ratio(DI), state of residence of the applicant (ST), status of home ownership – whether they rent, own or have a home mortgage (HO), the applicant's monthly income (MI), the credit score for the applicant provided in a range form (FICO), the number of open credit lines (OC), the total amount outstanding of all lines of credit (RC), the number of credit inquiries in the last 6 months(CI), and the applicant's length of employment at their current job (LE). There were 7 missing data points: 1 from monthly income, and 2 each from number of open credit lines, amount outstanding of all lines of credit and number of credit inquiries in the last 6 months. Also, there were 4 outliers of data where the monthly income was greater than $30,000, so these were removed from the data set.

An initial simple linear regression was performed on the formula $IR=b_0 + b_1(FICO) + e$ where $b_0$ is the intercept and $b_1$ represents the change in interest rate associated with a change of 1 unit of FICO score and e represents the error. An adjusted R-squared value of 0.5026 was reported for this model indicating approximately 50% of the variation of interest rate can be explained by the FICO score. A residual standard error of 2.947 was found with a p-value of less than $2x10^{-16}$ indicating the association is statistically significant as would be expected. Figure 1 is a plot of the interest rate data versus FICO score with a line indicating the simple linear fit. Also illustrated in figure 1 is the large impact the length of loan has on the interest rate, note that loans of length 60 months have, on average, higher interest rates for the same FICO score compared to loans of length 36 months. The effect of loan length on interest rate is further illustrated in figure 2, which is a boxplot of interest rate versus loan length.

Prepared by: Mary Anne Plano
January 14, 2014

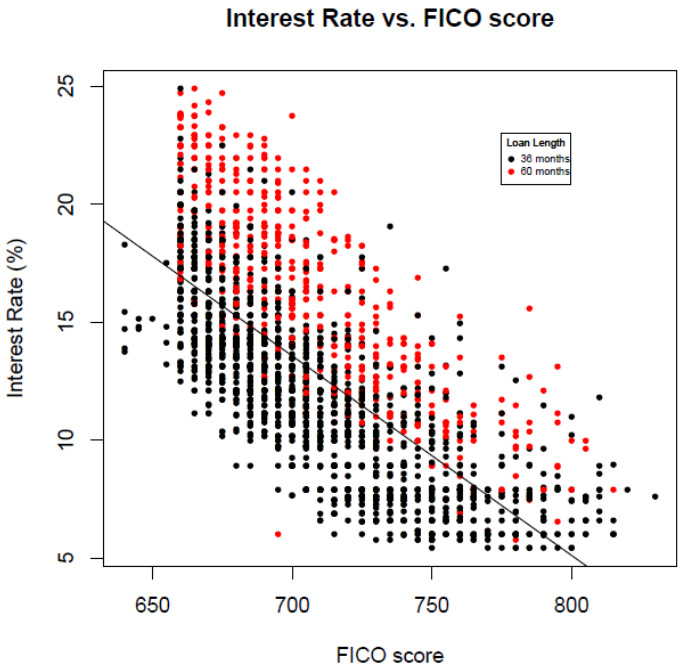


Figure 1 - A scatterplot of interest rate versus FICO score for all loan applicants. The line represents a linear fit with slope of -0.08 %/FICO score and intercept of 72.77 %. The effect of loan length on interest rate is illustrated through the use of color; 60 month loans (red circles) generally result in a larger interest rate than 36 month loans (black circles) for similar FICO scores.

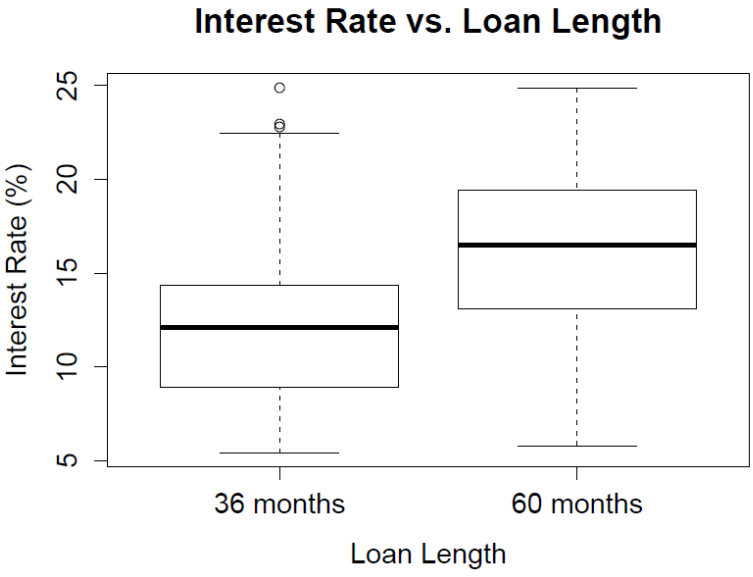


Figure 2 – A boxplot of interest rate versus loan length for all applicants. The median interest rate value for 36 months is 12.12%, while the median interest rate value for 60 months is 16.49%.

A multivariate linear model was developed for the interest rate. An initial model was tested using all given variables. Those variables that were deemed statistically insignificant by arbitrarily setting a p-value of 0.05 were removed one by one and tested versus the previous model using an analysis of variance. One variable – loan purpose – initially had a p-value greater than 0.05, but removing it from the model resulted in a statistically significant change based on the analysis of variance, so it was retained in the model. The scaled values of the slopes and intercept for the preliminary model can be expressed with the equation given below. Note, the variables are listed in decreasing order of the slope size.

$IR = B_0 + B_1(LL) + B_2(FICO) + B_3(ST) + B_4(AF) + B_5(LP) + B_6(CI) + B_7(AR) + B_8(HO)+B_9(OC)+B_{10}(MI)+e$

**Table of scaled coefficients for the preliminary multivariate linear regression model.**

| Coefficient | Value with error | Related factor |
|---|---|---|
| $B_0$ | 0.262 +/-0.164 | Intercept |
| $B_1$ | 0.746 +/-0.027 | LL-loan length |
| $B_2$ | -0.729 +/-0.010 | FICO – low end of FICO range |
| $B_3$ | Varies by state ~ -0.3 to -0.4 +/-0.15 to 0.20 | ST – state of residence |
| $B_4$ | 0.208 +/-0.041 | AF – amount funded |
| $B_5$ | 0.217 +/-0.115 (Moving)<br>0.146 +/- 0.078 (Other) | LP-loan purpose |
| $B_6$ | 0.116 +/-0.010 | CI - number of credit inquiries in the last 6 months |
| $B_7$ | 0.104 +/- 0.042 | AR – amount requested |
| $B_8$ | 0.048 +/- 0.023 (Rent) | HO-home ownership |
| $B_9$ | -0.039 +/- 0.010 | OC - number of open credit lines |
| $B_{10}$ | -0.032 +/- 0.012 | MI – monthly income |

The adjusted-R-squared value for this model is 0.765, indicating that 76.5% of the variance of interest rate can be explained with this model. The residual standard error for this model is 0.485 on 2426 degrees of freedom.

Four of the variables in the linear model have a negative slope; FICO, state of residence, number of open credit lines and monthly income. The negative slope indicates that as that variable increases, the interest rate decreases. The reason this occurs for the state of residence is that applicants from Arkansas happen to have the highest median interest rate and the dummy variable created for analyzing the effect of state is compared to Arkansas, since it occurs first alphabetically.

Three of the factor variables show a statistical significance only for some of the factor levels. For loan purpose, only “moving” and “other” are statistically significant factors in the model. For home ownership only the “rent” category is statistically significant and for the state

category most of the states are statistically significant – 40 out of 46. Those states which are not statistically significant are ones where few data points are provided, i.e. 10 or less. To further understand the impact of state of residence it would be good to analyze another data set containing a larger number of loans for each state considered.

It is clear that the variables utilized in the model could be interrelated; it seems especially likely that the FICO score and the amount funded may depend on the other variables. A linear model was applied to each of these, relating them to all the other variables. It was found that the FICO score is dependent on the debt to income ratio, number of inquiries in the last 6 months, the loan purpose and home ownership ("rent" condition). The amount funded was found to have a dependence on Amount Requested, Monthly Income, Revolving Credit Balance and number of inquiries in the last 6 months. So multiplied factors were added into the model and then those which did not have statistical significance were removed one by one until the final model was obtained. The final model can be expressed in the following equation:

$IR = B_0 + B_1(LL) + B_2(FICO) + B_3(St) + B_4(AF) + B_5(LP) + B_7(AR) + B_8(HO) + B_9(OC) + B_{10}(MI) + C_1(CI)*(FICO) + C_2(LP)*(FICO) + C_3(HO)*(FICO) + C_4(AF)*(MI) + C_5(AF)*(CI) + e$

**Table of scaled coefficients for the final multivariate linear regression model.**

| Coefficient | Value with error | Related factor |
|---|---|---|
| $B_0$ | -.011 +/- 0.165 | Intercept |
| $B_1$ | 0.756 +/-0.026 | LL-loan length |
| $B_2$ | -0.538 +/-0.055 | FICO – low end of FICO range |
| $B_3$ | Varies by state ~ -0.3 to -0.4 +/-0.15 to 0.20 | St – state of residence |
| $B_4$ | 0.142 +/-0.047 | AF – amount funded |
| $B_5$ | 3.677 +/- 2.05(Moving)<br>3.541 +/- 1.32(Other) | LP-loan purpose |
| $B_7$ | 0.110 +/- 0.041 | AR – amount requested |
| $B_8$ | 1.817 +/- 0.44 | HO-home ownership (rent) |
| $B_9$ | -0.043 +/- 0.010 | OC - number of open credit lines |
| $B_{10}$ | -0.066 +/- 0.020 | MI – monthly income |
| $C_1$ | $9.7X10^{-5}$ +/-$1.9x10^{-5}$ | (CI)*(FICO) – FICO times number of inquiries in last 6 months |
| $C_2$ | $-6x10^{-3}$ +/-2x10-3 | (LP)*(FICO) – FICO times loan purpose |
| $C_3$ | $-2.512x10^{-3}$ +/-$6x10^{-4}$ | (HO)*(FICO) - FICO times home ownership(rent) |
| $C_4$ | $7.191x10^{-10}$ +/-$3x10^{-10}$ | (AF)*(MI) – amount funded times monthly income |
| $C_5$ | $2.916x10^{-6}$ +/-$1x10^{-6}$ | (AF)*(CI) – amount funded times number of inquiries in last 6 months |

The addition of these multiplied terms slightly increased the adjusted-R-squared term to 0.775 and yielded a residual standard error of 0.474 on 2408 degrees of freedom. It was found that

after adding in these multiplied terms the individual term for number of credit inquiries in the last 6 months was no longer significant and so was deleted from the model. The higher degree of complexity added to the model by the multiplied factors may not be worth it as it yields only a 1% improvement in the adjusted–R-squared term. Also, the intercept term has degraded for this model: it is no longer a statistically significant value and the error is larger than the coefficient. However, these confounders may indicate a fruitful path for further analysis i.e. pursuing non-linear models using a combination of the interrelated variables.

## Conclusion

An analysis of 2500 peer-to-peer loans provided through the Lending Club indicates there is a significant association between interest rate and credit score, loan length, amount funded and in some cases the state of residence of the applicant. Other factors that played a role in the linear model were loan purpose, number of credit inquiries in the last 6 months, amount requested, home ownership (rent), number of open credit lines and monthly income. The higher the credit score, the lower the interest rate. The longer the loan or the more funded, the higher the interest rate. With a simple linear model based solely on single factors 76.5% of the variation in interest rate can be explained. A more complex model containing interactive terms is only able to account for an additional 1% of the variation. Additional research could be done to quantify the effect of state of residence on interest rate charged; a more complete data set would be needed. It might be possible to link the state of residence with other geographic-based economic factors.

## References

[1]Lending Club Website https://www.lendingclub.com/

[2]"The 5 Biggest Factors that Affect Your Credit" by Amy Fontinelle 9/10/10
http://www.investopedia.com/articles/pf/10/credit-score-factors.asp

[3] Source for loans data http://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

[4] R Core Team (2012). "R: A language and environment for statistical computing." URL:
http://www.R-project.org

[5] R script used for analysis -
https://github.com/maplano/projects/blob/master/script%20for%20loans%20data%20paper.txt

[6] “Using R for Linear Regression”, Montefiore Institute, University of Liege
http://www.montefiore.ulg.ac.be/~kvansteen/GBIO0009-1/ac20092010/Class8/Using%20R%20for%20linear%20regression.pdf

[7]”Getting Started in Linear Regression using R” by Oscar Torres-Reyna, Princeton
http://dss.princeton.edu/training/Regression101R.pdf

[8] “Multiple Regression Tutorial” by William B. King, Ph.D., Coastal Carolina University.
http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html

Prepared by: Mary Anne Plano
January 14, 2014