## INTRODUCTION

Lending Club is a financial institution that provides a medium for borrowers and investors to benefit from a low-cost online environment to meet their financial needs [1]. According to Lending Club, the better rates offered to applicants are influenced by several factors such as creditworthiness, debt to income ratio, and recent credit activity [2].

This study was conducted to investigate the various factors affecting the interest rates offered by Lending Club. It was found that there is a significant association between interest rate offered and the applicant's credit score, amount of loan requested, and the length of the loan. The analysis performed suggests that longer-term loans, higher loan amounts, and lower credit scores are associated with increased interest rates.

## METHODS

### *Data Collection*

The data used for this analysis was downloaded from Coursera.org on November 8, 2013 which includes a sample of 2,500 loans from Lending Club [3].

### *Exploratory Analysis*

Non-graphical methods such as data summaries and cross-tabulation, and graphical methods such as scatter plots and boxplots were utilized to perform exploratory data analysis in order to (1) Identify missing values, (2) Identify outliers and unusual features in the data, (3) Verify quality of the data, (4) Explore correlations between all variables, (5) Transform variables to the appropriate class for further analysis.

### *Statistical Modelling*

To exhibit the relationship between the interest rate and the explanatory variables, a multivariate linear regression model was employed [4]. The explanatory variables and the model were selected based on the outcome of our exploratory analysis and the coefficients in the model were estimated with ordinary least squares [5].

### *Reproducibility*

The current analyses are reproducible through accessing the R markdown file Ass1.Rmd [6]. To obtain the exact results given in the current document, the same data source must be used. The applicants' requirements set by Lending Club may change over time and hence one may not reach the same conclusions if the analyses are performed on a different dataset.

## RESULTS

The data used in this study contain various variables which hand in useful information about 2,500 loan applications. Table 1 provides a brief summary of these variables.

**Table 1. Summary of Loans Data [3]**

| Variable | Description | Unit | Class |
|---|---|---|---|
| **Amount Requested** | The amount requested in the loan application | $ | Integer |
| **Amount Funded by Investors** | the amount loaned to the individual | $ | Numeric |
| **Interest Rate** | The lending interest rate | % | Factor |
| **Loan Length** | The length of time of the loan | Months | Factor |
| **Loan Purpose** | The purpose of the loan as stated by the applicant | N/A | Factor |
| **Debt to Income Ratio** | The percentage of consumer's gross income that goes toward paying debts | % | Factor |
| **State** | The abbreviation for the U.S. state of residence of the loan applicant | N/A | Factor |
| **Home Ownership** | A variable indicating whether the applicant owns, rents, or has a mortgage on their home | N/A | Factor |
| **Monthly Income** | The monthly income of the applicant | $ | Numeric |
| **FICO Range** | A range indicating the applicant's FICO scores. | N/A | Factor |
| **Open Credit Lines** | The number of open lines of credit the applicant had at the time of the application | N/A | Integer |
| **Revolving Credit Balance** | the total amount outstanding all lines of credit | $ | Integer |
| **Inquiries in the Last Six 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 | N/A | Integer |
| **Employment Length** | Length of time emplyed at current job | Years | Factor |

As part of the data cleaning and screening process, interest rate, debt to income ratio, FICO range, and employment length were coerced to numeric variables. The upper value for FICO range was used. The lower value or the middle value could have been picked alternatively. Since the the FICO range intervals were not substantially wide, the selection of the FICO score is not an issue. The data points with employment length < 1 year , 10+ years, and n/a were replaced with 0, 10, and the average of all the other data points in employment length respectively. An odd data point in monthly income was observed ($102,750). This data point was assumed to be an extreme outlier since it's unlikely for someone with this monthly income to ask for a small loan. Therefore, the observation for that data

point was removed. A total of seven missing values were found which correspond to two observations. These observations were also omitted from the data since they do not affect our analysis significantly.

A correlation matrix and a pairwise scatter plot among the numeric variables demonstrated a strong negative correlation between interest rate and FICO score. Amount requested seemed to have some association with interest rate as well. Therefore, these two variables were perceived to be potential candidates to explain the variance in interest rate.

Looking at the correlation matrix, revolving credit balance, open credit lines, debt to income ratio, monthly income, inquires in the last six months, and employment length were not highly correlated with each other and other variables. However, a research on calculation of the FICO score revealed that all these variables are confounded with the FICO score [7]. As a result, these variables were considered to be redundant and hence were eliminated from the study. Including redundant variables may skew the model due to multicollinearity [8].

After examining numeric variables, a few boxplots were generated to examine the effect of the factor variables on interest rate. Except for Home Ownership, the graphs suggested that there may be a difference in the mean of the levels within Loan Length, Loan Purpose, and State groups. Analysis of variance was employed to see if this was the case. Looking at the values of etaSquared, it was seen that only loan length explained a decent amount of variance in interest rate. Based on the results of this analysis, state and loan purpose variables were also eliminated from the study.

After eliminating all the redundant variables and variables which did not significantly account for the variance in interest rate, the final regression model was constructed as follows:

$$I = b_0 + b_1(FICO) + b_2(Amount) + b_3(Length) + e \qquad (1)$$

Where $b_0$ is the intercept, $b_1$ represents the change in Interest Rate $I$ associated with a change of 1 unit in FICO Score, $b_2$ represents the change in Interest Rate $I$ associated with a change of 1 unit in Amount Requested, $b_3(Length)$ represents a factor model with 2 different levels for Loan Length, and $e$ represents everything that was not measured and is not explained by the model.

A highly statistically significant ($P=2e^{-16}$) association between interest rate and FICO score, amount requested, and loan length was observed. A change of 1 unit in FICO score corresponds to a drop of $b_1 = -0.087$ in percentage of interest rate while other variables are held constant (95% CI: -0.090,-0.085). A change of 1 unit in loan amount requested translates to an increase of $b_2 = 0.000138$ in percentage of interest rate while other variables are held constant (95% CI: 0.000126, 0.000150). Moving from 36 months to 60 months loan length corresponds to an increase of $b_3 = 3.29$ in the percentage of interest rate while other variables are held constant (95% CI:3.08 , 3.51).

Figure 1 clearly illustrates the results of the final regression model as described in equation (1). One can see that higher interest rates correspond to lower FICO scores, higher loan amounts, and longer-term loans.

## CONCLUSIONS

In conclusion, our analysis suggests that there exists a significant negative association between the interest rate and applicant's FICO score. Moreover, a significant positive association among interest rate, amount of loan requested, and length of the loan was observed. This analysis estimated the relationship between interest rate and the explanatory variables using a linear multivariate regression model. Throughout the study, a number of confounding variables were examined. Revolving credit balance, length of employment, monthly income, and number of credit lines were some of these variables which were eliminated to improve the final regression model.

The results of this study are in complete agreement with domain knowledge. As known, the lower the FICO score and the higher the loan amount and the longer the loan term, the higher the risk of default. Hence to compensate for the added risk, a higher interest rate is charged.

Although the final model explains a substantial amount of variance in the dependent variable ($R^2$=0.75), there are still some other factors influencing the interest rate that were not revealed by the model. The assumption of normality of the independent variables was ignored in the preliminary exploratory analysis of the data. Even though the independent variables are not highly skewed, a transformation of the variables may further improve the model. Additional assumptions such as homoscedasticity and autocorrelation as well as the effect of the extreme outliers and influential data points were not considered due to the narrow scope of knowledge of the author [9]. Future studies with more rigorous diagnostic tests on the data could reveal potential issues with the results of the current study and lead to a more legitimate regression model explaining the relationship between the variables.

## REFERENCES


[1] Lending Club website. URL: https://www.lendingclub.com/public/about-us.action. Accessed 11/8/2013

[2] Lending Club Website. URL: https://www.lendingclub.com/public/how-we-set-interest-rates.action. Accessed 11/8/2013

[3] Coursera Data Analysis course webpage. URL: https://class.coursera.org/dataanalysis-002/human_grading/view/courses/971332/assessments/4/submissions. Accessed 11/8/2013

[4] Coursera Data Analysis course notes. URL: https://d396qusza40orc.cloudfront.net/dataanalysis/multipleVariables.pdf. Accessed 11/8/2013

[5] Wikipedia webpage. URL: http://en.wikipedia.org/wiki/Ordinary_least_squares. Accessed 11/8/2013

[6] R markdown webpage. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 11/8/2013

[7] myfico.com webpage: URL: http://www.myfico.com/crediteducation/whatsinyourscore.aspx. Accessed 11/8/2013

[8] Wikipedia webpage. URL: http://en.wikipedia.org/wiki/Multicollinearity. Accessed 11/8/2013

[9] University of Delaware, Public Management and Applied Statistics course notes: http://www.udel.edu/htr/Statistics/Notes816/class16.PDF. Accessed 11/8/2013