**Title: Lending Interest Rate is associated with FICO score, loan length and number of earlier loan inquiries.**

**Introduction:**

Lending Club opened in 2007 with one simple mission: create an alternative to banks that offers both borrowers and investors a great rate. Lending Club allows its members to directly invest in and borrow from each other, thus avoiding the cost and complexity of the banking system and pass the savings on to both investors and borrowers. Both sides can win: better rates to borrowers and better returns to investors.

The variable that plays most important role in determining interest rate is borrower FICO score [1] which is a measure of borrower creditworthiness.

The purpose of this analysis is to determine whether there are other factors contributing to the loan interest rate after taking into account the applicant's FICO score.

**Methods:**

*Data Collection*

For our analysis we used the data consisting of sample of 2,500 peer-to-peer loans issued through the Lending Club [2]. The data were downloaded from Amazon Web Services hosting service [3] on 8th February 2013 using the R programming language [4].

*Exploratory Analysis*

Exploratory analysis was performed by examining tables and plots of the observed data. We identified transformations to perform on the raw data on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the variables having statistically significant correlation with loan interest rate that would be used in creating statistical model.

During exploratory analysis, besides FICO score, loan duration and number of applicant loan inquiries in the last 6 months were found to be covariates having statistically important correlation with loan interest rate.

*Statistical Modeling*

To relate loan interest rate to FICO score, loan duration and number of applicant loan inquiries in the last 6 months we performed a standard multivariate linear regression modeling [4]. Model selection was performed on the basis of our exploratory analysis. Coefficients were estimated with ordinary least squares and standard errors were calculated using standard asymptotic approximations [5].

**Results:**

The loans data used in this analysis contains information on the loan interest rate (Interest.Rate) and following variables:

Amount.Requested

Amount.Funded.By.Investors

Loan.length

Loan.Purpose

Debt.to.Income.Ratio

State

Home.ownsership

Monthly.income

FICO.range

Open.CREDIT.Lines

Revolving.CREDIT.Balance

Inquiries.in.the.Last.6.Months

Employment.Length

These variables are described in detail in the dataset codebook [5]. We identified two observations with missing values in the data set and these observations were excluded from further analysis. During exploratory analysis one of the observations was found to be outside of linear relationship between interest rate (Interest.Rate) and Inquiries.in.the.Last.6.Months and this observation was excluded for final linear modeling.

Most earthquakes had a small (less than 3 on the Richter scale - 85% of quakes) or medium (3-5 on the Richter scale - 11% of quakes) magnitude. The distribution of earthquake depths was heavily right skewed. Based on the distribution of the earthquake depths we recognized that a transform was necessary to improve the performance of linear regression techniques; we performed a log base 10 transform of the earthquake depths. Subsequent analyses focus on this transformed depth variable.

We first fit a regression model relating earthquake magnitude to earthquake depth. The residuals showed patterns of non-random variation. We attempted to explain those patterns by fitting models including potential confounders. Our final regression model was:

EM = b0 + b1 log10(ED) + f(Lat) + g(Lon) + h(NST) + e

where b0 is an intercept term and b1 represents the change in earthquake magnitude on the Richter scale associated with a change of 1 unit in log base 10 kilometers for earthquakes at the same latitude, longitude, and measured by the same number of stations. The terms f(Lat), g(Lon), and h(NST) represent factor models with 5 different levels each for latitude, longitude, and number of measurement sites. The error term e represents all sources of unmeasured and unmodeled random variation in earthquake magnitude. Our final regression model appeared to remove most of the non-random patterns of variation in the residuals.

We observed a highly statistically significant (P = 8e-15) association between earthquake magnitude and earthquake depth. A change of one unit in log base 10 kilometers of depth corresponded to a change of b1 = 0.41 on the Richter scale (95% Confidence Interval: 0.31, 0.51). So for example, for two earthquakes at the same latitude, longitude, and measured at the same number of stations, we would expect an earthquake measured at a depth of 10 kilometers to be 0.57 units higher on the Richter scale than one measured at a depth of 1 kilometer. We would expect the same difference between earthquakes measured at depths of 10 and 100 kilometers and so forth.

**Conclusions:**

# References

|  |  |
| --- | --- |
| [1] | W. community, "Credit score in the United States," Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Credit\_score\_in\_the\_United\_States. [Accessed 17 February 2013]. |
| [2] | “Lending Club Page,” [Online]. Available: https://www.lendingclug.com/home.action. |
| [3] | "Data on Lending Club Loans," [Online]. Available: https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. [Accessed 8 February 2013]. |
| [4] | R Core Team, "R: A language and environment for statistical computing," 2012. [Online]. Available: http://www.R-project.org. [Accessed 8 February 2013]. |
| [5] | "Codebook for the available dataset," [Online]. Available: https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf. [Accessed 8 February 2013]. |