Introduction

The Lending Club is a company that allows people get loans from a lot of investors and investors to borrow money in what is considered a peer-to-peer loan. The basic premise is: Borrowers apply for money, get funded and finally repay. Investors open an account, build a portfolio and invest money [1]. The interest rate of these loans is determined by the Lending Club on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and credit worthiness scores.

Understanding how the interest rate is calculated can help borrowers to get better deals as an lower interest rate is always better. Here we performed an analysis to determine if there was a significant association between some of the variables previously mentioned such as the FICO rate and loan length and the final Interest Rate of the loan. The analysis made suggests that there is a relation between the final Interest rate and some variables such as the borrower FICO Range and the length of the loan. While there is little relation between the purpose of the loan and the interest rate applied/

Methods

Data collection

For our analysis we used a sample of 2,500 peer-to-peer loans issued through the Lending Club each record has the following information: amount requested by the borrower, amount funded by the investors, loan length, loan purpose, debt to income ratio, state of the borrower, homeownership , monthly income, FICO range, open credit lines, revolving credit balance, inquiries in the last 6 months, employment length, and the final interest rate of the loan.

Exploratory analysis

Using tables and graphics it was possible to identify missing values, quality of the data, necessary transformations to be done and to discover some of the relations between the target variables (interest rate) and its inputs.

Data transformation

The data followed some of the basic rules of what is considered tidy data but it was necessary to transform some of the columns in order to make the analysis possible. The main changes was to transform columns that represent numerical values but because of the format of the data does not represent number on the computer. Those columns where: interest rate, loan length, debt to income, FICO range, employment length. For the variables with percent values only the percent symbols was removed and the number was leaved, so instead of going from 0 to 1 the values are from 0 to 100. For the FICO Range variable only the lower limit was used for the

analysis.

Statistical Modeling

To relate the Interest rate to the other variables a standard multivariate linear regression was performed. The variables selection were based on the correlation between the variable and the target variable (Interest Rate). Multiple regressions were performed before taking the final decision.

Reproducibility

Because of security concerns it is not possible to share the R markdown file but the code is reproducible using the R markdown format and using the R version 2.14.

Results

It was possible to identify a few missing values in some variables specially in the Employment Length. Those missing values were imputed using the KNN clustering method, since the missing values where a small portion of the data the changes in the final results were minimal.

Using the singular value decomposition method it was possible to observe that there is at least one pattern in the data that can explain the variance. The first singular value <<check notation>> explains 82.24% of the variance in the data and the second singular value explains 16.13% of the variance. Together both singular vectors explains 96.37% of the variance in the data. [Figure 1 - Left Panel]

The distribution of Interest Rate showed that is not very skewed (0.279) so is not very necessary to do any sort of transformation to remove the skewness. On the other hand some variables needed to be transform as they have a higher skewness specially Monthly Income and Revolving Credit Balance with values of 8.469 and 5.38.

Intuition tell that there is a relation between the Interest Rate and some variables such as FICO Range and Amount Requested while there is possible very little relation between the Interest Rate and the State. To check that a relation the variables the correlation between the Interest Rate was calculates the most correlated variables are: FICO Range (-0.709111) and Loan Length (0.423694).

To test the significance of the the relation between variables a linear regression was fitted using most correlated variables. The equation for the regression was:

$$Interest\ Rate\ =\ b0 + b1(FICORange)\ +\ b2(LoanLength) + e$$

The statistical significance founded was very high for FICO range and for Loan Length which is

the lowest possible the computer where the tests were runned can found, that is: 2e-16. Using other software was possible to find a significance of 4.27e-258 for Loan Length.

The values of b1 and b2 are -0.0852 and 0.1819 with a 95% interval confidence of (-0.088, -0.083) and (0.173, 0.191) respectively. This tell us that with two persons with the same FICO range a change in the Loan Length of 1 Months will increase the Interest Rate 0.1819 (Interest rate goes from 0 to 100). The R squared of the regression was 69%. This tell us that 69% of the interest rate is explained by those two variables.

With respect to some categorical variables it was found that some loan purposes have a significant statistical relation [Figure 1 - Right Panel]. For example a purpose of debt consolidation has a significance of 3.05e-05 while other purposes do not have a relevant statistical significance such a purpose of medical with a significance of 0.572. For this reason was not considered as a confounder in this analysis.

Conclusions

This analysis suggests that there is a significant relation between the FICO Range of a person and the Loan Length and the Interest Rate applied to a loan of the Lending Club. Those two variables explain almost 70% of the interest rate applied to a loan. While other variables are statistically significant and can be used as confounders do not represent a considerable change in the final interest rate.

This analysis can help people who are currently thinking on getting a loan or borrow some money with the Lending Club. Borrowers should pursue lower loan lengths and higher FICO range while lenders should look for lower FICO Ranges and higher Loan lengths in order to maximize the interest rate with the risks this entails.

References

[1] The Lending Club, how it works: URL: https://www.lendingclub.com/public/how-peer-lending-works.action Accecced: Feb 15, 013.