**Analytics Essentials Project**

**PEER-TO-PEER Loan Data Analysis**

1. **INTRODUCTION**

Peer-to-peer lending [1] is the practice of lending money to unrelated individuals, or "peers", without going through a traditional financial intermediary such as a bank or other traditional financial institution and is conducted for profit. Lending Club [2] is one of such a peer to peer lending company located in US.[3]Lending Club’s interest rates take into account credit risk and market volatility. This adjustment of risk and volatility is designed to cover expected losses. A grade is determined for a loan based on the borrower’s credit score, and combination of several indicators of credit risk from the borrower’s credit report and loan application.

The purpose of this analysis is to identify and quantify associations between the interest rate of the loan and the other variables in the data set. In particular, to figure out any of these variables have an important association with interest rate after taking into account the applicant's FICO score. For example, if two people have the same FICO score, can the other variables explain a difference in interest rate between them?

1. **METHODS**
   1. ***Data collection***

The data consist of a sample of 2,500 peer-to-peer loans issued through the Lending Club[[1](https://www.lendingclub.com/home.action)].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 creditworthiness scores.

Data set was obtained from the location hosted on the amazon web services location

<https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv> in the name of loansData.csv.

The associated code book to interpret the variables is found here:

<https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf>

* 1. ***Data Summary***

Data set contain 2500 observations across 15 columns.

***Dependent variable***

1. Interest Rate (in %)

***Independent Variables***

1. Loan Purpose
2. Home Ownership
3. State
4. Employment Length
5. FICO Credit Score Range
6. Loan Amount Requested (in USD)
7. Amount Funded by Investors (in USD)
8. Loan Length (in months)
9. Debt to Income Ratio (in % )
10. Monthly Income (in USD)
11. Number of Open Credit Lines
12. Revolving Credit Balance
13. Loan Inquiries in the Last 6 Months
    1. ***Data Transformation***

A quick summary was done on the 2500 observations to get an overall picture of the data. Based on the summary, below are the activities performed.

* There were a total of **7 missing values** and **77 values with "n/a"** in the data set. **These rows were removed** from the data set for the purpose of analysis.
* FICO Credit score was given in ranges from 640 to 834 with an interval of 4. So **FICO score ranges were converted to a numeric value** taking into account the higher value of a given range.
* **“%” sign was removed** from Interest rate and debt to income ratio values
* Characters/Signs like **“years”,”<”,”>” were removed** from the Employment Length values.
* Values with “n/a” was converted into “NA” and omitted from the dataset.
* **Names of all the variables were converted to smaller caps** for easier programming and interpretation.
  1. ***Exploratory Analysis***

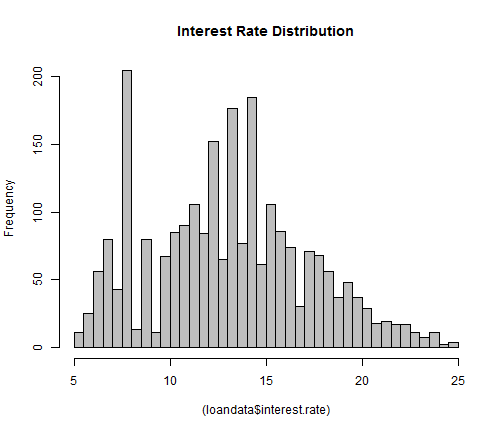
A histogram on Frequency distribution of Interest rate reveals the distribution is normal but on further investigation by introducing break of 50 shows that the distribution is not normal **Fig.(1)**. To go further and establish this statistically, Shapiro-Wilk test [7] of normality is done on the dependent variable. The test confirms that interest rate distribution is not normal and with a p-value closer to 0, it’s wildly unlikely that the distribution is normal.

Shapiro-Wilk normality test

data: loandata$interest.rate

W = 0.9815, p-value < 2.2e-16

p-value<0.05. Reject H0, so the 'Y' is not normal



**Fig.(1)**

Later interest rate was checked against the most likely independent variable the loan length through multiple visuals of histogram, density plot and there seemed to be a difference in the distribution when the loan length factor was considered. The box plot clearly suggested the means are different in each case but there seemed to be an overlapping.

To establish this significantly, hypothesis testing was performed. Since we wanted to compare the means of two groups and establish statistically that they are different two sample t-test was chosen. Since it mandates the Y to be normally distributed, I went ahead with non-parametric equivalent Mann-Whitney test a.k.a two-sample Wilcoxon test [8]. A p value of <0.05 confirmed that the median of two groups are different and the distribution isn’t similar.

Wilcoxon rank sum test with continuity correction

Data: interest.rate by loan.length

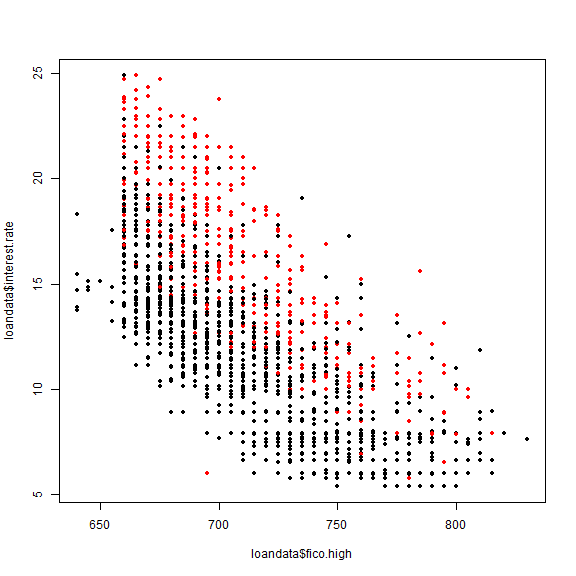
W = 231926, p-value < 2.2e-16

Alternative hypothesis: true location shift is not equal to 0

**Null Hypothesis H0:** Distribution of Two groups are identical or equal median

**Alternate Hypo H1:** Distribution of two groups are not identical

With a p-value<0.05, we reject null hypothesis and conclude the distribution of interest rate for 36 month period and 60 month period are not equal



Multiple scatter plots were tried to weigh in their effect on Interest rate v/s fico rate but, none of the factors other than loan length showed us a significant association/pattern. Out of intuition monthly income was taken into account and created new ranges with 5 groups to weigh in if at all it made any impact on the interest rate. But it did not reveal any significant pattern in the visual.

A scatter plot matrix was created to see the association between numeric variables with the loan length factor considered. It revealed the association of Amount funded by investors and Amount Requested with interest rate. This was statistically confirmed with a correlation matrix.

This shows how FICO Score is negatively correlated with Interest Rate, explaining higher the score lower the interest rate.

**Correlation Coefficient Table**

|  |  |
| --- | --- |
| **X's** | **Interest Rate** |
| Amount funded by investors | 0.33254127 |
| Amount requested | 0.327497515 |
| Debt to income ratio | 0.174868728 |
| Employment length | 0.051890066 |
| Fico high | -0.70753788 |
| Inquiries in the last 6 months | 0.161909773 |
| Interest rate | 1 |
| Monthly income | 0.009292971 |
| Open credit lines | 0.092543286 |
| Revolving credit balance | 0.067289914 |

Correlation matrix also revealed a strong correlation between two of the independent variables, creating a case of multicollinearity.

|  |  |
| --- | --- |
| **X’s** | amount.requested |
| amount.funded.by.investors | 0.96875 |

* 1. ***Statistical Modeling***

To statistically establish the multicollinearity with in the related independent variable a simple model was created and the Variable Inflation Factor was checked for the two variables. Going by the thumb rule of variables with VIF>5 exhibits multicollinearity we confirm the fact that amount funded and amount requested has multicollinearity.

> rmvif <- lm(interest.rate~amount.funded.by.investors+amount.requested+fico.high, data=loandata)

> vif(rmvif)

amount.funded.by.investors amount.requested

**16.268 16.293**

fico.high

1.009

Since both the variables were equally correlated with Interest rate, it was decided not to drop them but instead go for a dimension reduction technique, i.e. principal component analysis. For this both the above variables were standardized and PCD was done. From the summary we could figure out that the Component 1 variable itself cumulatively explained 98% of variation in interest rate. Based on this Comp.1 model was selected for modelling.

> summary(.PC) # proportions of variance

Importance of components:

Comp.1 Comp.2

Standard deviation 1.4031 0.17677

Proportion of Variance 0.9844 0.01562

Cumulative Proportion 0.9844 1.00000

So a multivariate linear model was developed with independent variables FICO score, Loan Length, PC1 (which is a new variable created by weighted importance between amount requested and amount funded)

> LinearModel.4 <- lm(interest.rate ~ fico.high + as.factor(loan.length) + PC1, data=loandata)

> summary (LinearModel.4)

Call:

lm(formula = interest.rate ~ fico.high + as.factor(loan.length) +

PC1, data = loandata)

Residuals:

Min 1Q Median 3Q Max

-9.761 -1.447 -0.125 1.270 10.281

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 74.06693 0.86691 85.4 <2e-16 \*\*\*

fico.high -0.08744 0.00123 -71.3 <2e-16 \*\*\*

as.factor(loan.length)[T.60] 3.29025 0.11256 29.2 <2e-16 \*\*\*

PC1 0.77073 0.03344 23.1 <2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.1 on 2417 degrees of freedom

Multiple R-squared: 0.746, Adjusted R-squared: 0.745

F-statistic: 2.36e+03 on 3 and 2417 DF, p-value: <2e-16

The model gives an adjusted R-Square value of 0.745. Implying 74% of the variation in interest rate is explained by this model.

1. **RESULTS**

A multivariate linear regression model was developed to better explain the interest rate change given a FICO Range.

The model developed explaines 74% of the variablitiy in interest rate. Since loan length was taken as a factor while modelling, it becomes a dummy variable in the final model with loan length=36 as reference scale. So for the given fico score if the amount funded by investors is constant, interest rate on a 60 month loan term would have β3=3.29 times higher that than that of 36 month term one.

***Interest rate= β1 + β2\* Fico.high + β3\*loan.length[=60) + β4\*PC1 +e***

|  |  |  |
| --- | --- | --- |
| **Coefficients:** |  |  |
| **(Intercept)** | β1 | 74.06693 |
| **fico.high** | β2 | -0.08744 |
| **as.factor(loan.length)[T.60]** | β3 | 3.29025 |
| **PC1** | β4 | 0.77073 |

The model developed explaines 74% of the variablitiy in interest rate. Since loan length was taken as a factor while modelling, it becomes a dummy variable in the final model with loan length=36 as reference scale. PC1 could be replaced by **(amount.funded.by.investors-Meana.f.b.i)/Standard deviationa.f.b.i.**

1. **CONCLUSION**

As a result of the analysis, the FICO.Range,Loan.Length,Amount.Requested, Amount.Funded.By.Investors were found to have a significant association with Interest rate. Multi-variate model gives a more accurate interest value given a FICO range.

1. **RECOMMENDATION**

Even though the goodness of fit for this model appear to be good interms of R-Squared value, analyzing the residuals reveals that it is not normal. So there is further scope for the improvement of the model with respect taking lot to the base of 10. Another problem with the interest rate distribution which seems to be skewed could be analysed to see if at all it could be made normal so as to explain the variations better throught the model.

1. **REFERENCES**

[1] <https://www.lendingclub.com/home.action>

[2] <http://en.wikipedia.org/wiki/Peer_to_peer_lending>

[2] <http://en.wikipedia.org/wiki/Lending_Club>

[3] <https://www.lendingclub.com/public/how-we-set-interest-rates.action>

[4] <http://en.wikipedia.org/wiki/P-value>

[5] <http://en.wikipedia.org/wiki/Simple_linear_regression>

[6] <http://blog.yhathq.com/posts/r-lm-summary.html>

[7] <http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test>

[8] <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/wilcox.test.html>