Data Analysis –

Identify and Quantify the relation of interest rate with other variable factors at Lending Club

---

"*Data Analysis*" Project 1

Course Offering of

JOHNS HOPKINS UNIVERSITY and COURSERA

Prepared By Bikal Basnet, basnet.beekal@gmail.com

Table of Contents

Introduction:

Lending club is an online financial community that brings together the creditworthy borrowers and savvy investors. [1] The platform benefits both of them financially, by providing smaller interest rates on loans while providing maximum returns to the investors, all by keeping the operational costs low. [1]

Interest rates are one of the major concerns for the borrowers. And borrowers want to have as lower interest rate as possible However; different financial institutions take into account different characteristics to determine the interest rate for borrowers.

Understanding the factors that affect the interest rate helps the potential would be borrowers to have lesser interest rates on their future loans. While for the lenders, it would be an opportunity to increase their loans volume or loan customers, as the lesser interest rates would likely make the loan lending club more appealing to its customers.

Here we performed an analysis to explore the factors that affect the interest rate(see fig a,b to observe the variables included in the linear model and the relationship between the residual error, interest rate and the fitted value), taking into consideration the FICO scores. Using exploratory analysis and standard multiple regression techniques we show that following variables as loan length (Val = 60 months), amount received, inquiries in the last 6 months, home ownership (Val = rent), state (Val = IL, LN, NJ, PA) have very significant relationship to interest rate ($P < 0.001$). Similarly state(Val = CA, CO, DC, FL, GA, MA, MI,MT, NY, OH, OK, VA,WA) show strong significant relationship ($P < 0.01$), while loan purpose (Val = moving, other), home ownership (Val = rent), state(Val = AL,AZ, AR, CT, DE, IA, KS, KY, LA, MD, MN, Mt, NC, NH, NV, RI, SC, TX, UT, WI, WV) show significant relationship ($P < 0.05$) with interest rate, even after adjusting for very significant factor fico score.

Methods:

Data Collection

For our analysis we used the data set sample of 2500 peer to peer loans issued by the lending club from the site https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. The data were downloaded on November 12, 2013 using the R programming language [2].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. Exploratory analysis was used to

(1) Identify missing values i.e.

(2) Verify the quality of the data

(3) Determine the terms to be used in the regression model

Statistical Modeling

To explore relationship between interest rate and other variables considering fico score, we performed a standard multivariate linear regression model [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].

Reproducibility

All analyses performed in this manuscript are reproduced in the R markdown file interestRateAnalysisFinal.Rmd [6]. The data was downloaded on 12 November 2013.

Analysis:

The lending club data used in this analysis contains information on the loan amount requested (Loan. Requested), amount invested by the borrower (Amount.funded.By.Investors), interest rate (Interest.Rate), and duration of the loan period (loan. Length), purpose of the loan (loan. Purpose), debt to income ratio (debt.to.income.ratio), state where the borrower resides (state), home status of the borrower whether he owns the house, rents it or has mortgage (home. Ownership), monthly income (monthly. Income), range indicating where the fico score of the borrower falls (fico.range), number of credits a person has (open.credit.lines), balance owed by the borrower to the diff creditors (revolving.credit.balance), number of inquiries made for the fico score in the last 6 months(inquiries.in.the.last.6.moths) , total duration the borrower has been employed(employment. Length). We identified few inconsistencies in the data set. They are tabulated below with the action taken to solve it and its justification. .

Problems	**Actions performed**	**Justification**
1. Missing Values	Only complete cases considered	Considering only the complete cases resulted in the loss of 5 out of 2500 observations. 5 observation loss is almost insignificant

2. Missing values in employment length	Considered missing values as the values of lowest order	The other option was to either remove all the 77 observations completely. Keeping them with lower order was more plausible

Quality of data

Amount funded by investors has values less than 0	The values less than 0 were set to 0	Amount cannot be less than 0 and hence the inconsistencies are considered typo errors and replaced with 0
Interest rate is character class (in format 12 %)	Removed percentage and then converted the class into numeric class	Interest rate is best suited as a quantitative numeric variable. Makes more sense and easy to deal during further exploration and modeling
Employment length is factored variable	Transformed into ordered factor with “na” as smallest Val and then converted to numeric class.	Addition, subtraction makes sense on employment length .More appropriate to inspect and model it as quantitative variable.
Debt to income rations in improper class (character) and format (12.5%)	Remove percentage char and transform to numeric character	Addition, subtraction makes sense on employment length .More appropriate to inspect and model it as quantitative variable.
Outliers detected on monthly income (65k,	Removed outliers	The values are outliers and lie well beyond the normal monthly income

102k)		distribution
Fico range is class variable with format (645-649)	Transformed as numeric class with lower end range value as its value	Subtractions on fico range make sense. Fico as quant variable makes more sense

All the variables were plotted against the interest rate. For variable with skewed data distribution, logarithmic operation was performed. The plot was observed, plotted with linear model fitted lines wherever necessary. The summary of the linear model and the linear model coefficients were observed. And based upon the plots and the linear observation, the following were the terms considered and rejected

Variable Factor taken into consideration for Data analysis

++

	T Val	Estimate	p
1. Loan. Length	23.34	4.27204	<2e-16
2. Amount.funded.by.investors	17.95	0.00018	"
3. amount, requested	14.85	0.00017	2e-16
4. Inquiries in the last.6.moths	8.398	0.56332	<2e-16
6. Debt.to.income.ration	8.6	0.095	"
7. Loan. Purpose			
8. Home. Ownership	Fval = 5.87 yes		
9. State	Fval = 1.21		

10. Fico.low

Variable Factors rejected

1. Open.credit.lines no relation observed during exploratory graph

2. Revolving.credit.balance 3.13 0.000015 0.0018 (Rejected: insignificant impact)

4. Employement.length no

++

The accepted terms were then checked for confounding values. The correlation computation pointed out a very highly confounded relation between, amount requested and amount funded by investors. The amount funded by investors was hence added to the rejected terms list, then.

Results

We first fit a regression model relating interest rate to the accepted variables taking into account the fico score. Our final regression model was:

Interest. Rate ~ b0 + a * loan. Length + b* scale (amount. Requested) + c* scale (inquiries.in.the.last.6.months) + d * scale (debt.to.income.ratio) + e * loan. Purpose + f * home. Ownership + g * state + h * scale (fico.low) + err

We observed the strong association between the interest rate and the following variables.

We observed very significant association (P = 2e-15) association between interest rate and

```
Loan.length60 months                    3.19325    0.11115  28.730  < 2e-16 ***
Scale (amount. Requested)                 1.17872    0.04823  24.441  < 2e-16
***
Scale (inquiries.in.the.last.6.months)  0.43364    0.04238  10.231  < 2e-16
```

Similarly we observed strongly significant association (P <0.01) between interest rate and

```
StateCA                                -1.95116    0.62578  -3.118 0.001842 **
State CO                                -1.81136    0.67161  -2.697 0.007044 **
StateVA                                -1.98401    0.66012  -3.006 0.002678 **
```

StateWA	-1.94081	0.67379	-2.880	0.004006 **
StateNC	-2.05698	0.66950	-3.072	0.002147 **
stateDC	-2.50936	0.87305	-2.874	0.004085 **
stateFL	-1.86692	0.63778	-2.927	0.003452 **
stateGA	-1.87882	0.65259	-2.879	0.004024 **

Similaryl we observed significant association between interest rate and

loan.purposemoving	0.98184	0.48346	2.031	0.042377 *
loan.purposeother	0.68294	0.32759	2.085	0.037197 *
home.ownershipOWN	0.39291	0.16070	2.445	0.014556 *
stateAL	-1.71335	0.70187	-2.441	0.014713 *
stateAR	-1.95653	0.83943	-2.331	0.019846 *
stateAZ	-1.41634	0.68787	-2.059	0.039597 *
stateCT	-1.53947	0.68231	-2.256	0.024143 *
stateDE	-2.27546	0.95157	-2.391	0.016866 *
stateIA	-5.31581	2.47979	-2.144	0.032160 *
stateKS	-1.49552	0.76256	-1.961	0.049974 *
stateKY	-1.51134	0.75238	-2.009	0.044674 *
stateLA	-1.50292	0.75673	-1.986	0.047138 *
stateMN	-1.80858	0.70219	-2.576	0.010065 *
stateMT	-2.10644	0.99294	-2.121	0.033988 *
stateNH	-1.85918	0.81493	-2.281	0.022612 *
stateNV	-1.84773	0.71742	-2.576	0.010067 *
stateOR	-1.23705	0.72364	-1.709	0.087488 .
stateRI	-1.93088	0.81287	-2.375	0.017607 *
stateSC	-1.76449	0.73003	-2.417	0.015722 *
stateTX	-1.31935	0.63758	-2.069	0.038621 *
stateUT	-1.75279	0.80249	-2.184	0.029044 *
stateWI	-1.73868	0.73686	-2.360	0.018374 *
stateWV	-1.65267	0.81412	-2.030	0.042466 *

Conclusions:

Our analysis suggests that there is a significant, positive association between interest rate and other variables considering the FICO score *(see Figure a, b)* such as
Loan length (Val = 60 months), amount received, inquiries in the last 6 months, home ownership (Val = rent), state (Val = IL, LN, NJ, PA) have very significant relationship to interest rate ($P < 0.001$). Similarly state(Val = CA, CO, DC, FL, GA, MA, MI,MT, NY, OH, OK, VA,WA) show strong significant relationship ($P < 0.01$), while loan purpose (Val = moving, other), home ownership (Val = rent), state(Val = AL,AZ, AR, CT, DE, IA, KS, KY, LA, MD, MN, Mt, NC, NH, NV, RI, SC, TX, UT, WI, WV) show significant relationship ($P < 0.05$) with interest rate, even after adjusting for very significant factor fico score. While most of the variables were observed to be positively associated, the state variable was observed to be negatively associated i.e. the states mentioned above pointed out the decrease in interest rate in those states. It is good aspect for the borrowers, if they happened to be from the state.

While our analysis is an interesting first step it is based on a limited sample of loan observations. A larger collection of representative sample may be more appropriate for understanding the relationship between interest rate and the variables. Our analysis may be of interest to borrowers seeking to better understand the lending process.

Bibliography

[1] l. club, "About us- Lending club," 2013. [Online]. Available: https://www.lendingclub.com/public/about-us.action. [Accessed 17 Novemeber 2013].

[2] R. C. T. 2012, "R Language and environment for statistical computing," [Online]. Available: www.R-project.org. [Accessed 12 November 2013].