Title: Peer-to-peer lending interest rate analysis

Introduction

“Peer-to-peer lending 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. This lending takes place online on peer-to-peer lending companies' websites using various different lending platforms and credit checking tools” [1].
Lending Club is one of the online financial community that brings together creditworthy borrowers and savvy investors replacing the high cost and complexity of bank lending with a faster way to borrow and invest [2].
Borrowers can apply for a loan online and get an instant rate quote. Lending Club claims that the interest rate of these loans is determined on the basis of characteristics of the person asking for the loan such as their employment history, credit history, and creditworthiness scores.
The aim of this assignment is to parse an analysis to determine if there is a significant association between the interest rate of the loan and the features of borrowers.
Using exploratory analysis and standard multiple regression techniques we show that there is a significant relationship between interest rate and FICO score, even after adjusting for confounding factors such as the amount funded, the loan length, open credit lines and Inquires on the last 6 months.
My analysis suggests that lower loan rate is associated with higher FICO score .

Methods:

Data Collection

For our analysis we used a sample of 2,500 peer-to-peer loans issued through the Lending Club. The data were downloaded, using the R programming language, from following link:
https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv
The code book for the variables in the data set is available here
https://spark-public.s3.amazonaws.com/dataanalysis/loansCodebook.pdf

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. I 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 terms used in the regression model relating interest rates and other variables like FICO scores, Debt-to-Income ratio, Open Credit Lines, Employment Length, Monthly Income and the ratio between Revolving Credit and the Monthly Income.

Statistical Modeling

To relate interest rate to scores, Debt-to-Income ratio, Open Credit Lines, Employment Length, Monthly Income and the ratio between Revolving Credit and the Monthly Income and FICO score I performed a standard multivariate linear regression model [4][5]. Model selection was performed on the basis of our exploratory analysis and prior knowledge of the relationship between [5].

Reproducibility

All analyses performed in this manuscript are reproduced in the R markdown file LoansDataAss1.Rmd [6]. To reproduce the exact results presented in this manuscript the cached version of the analysis must be performed, as the data available from LoansData.csv .

Results:

The peer-to-peer loans data used in this analysis contains information on the source network that report following information :

```
'data.frame':         2500 obs. of  14 variables:
 $ Amount.Requested              : int  20000 19200 35000 10000 12000 6000 10000 33500 14675 7000 ...
 $ Amount.Funded.By.Investors    : num  20000 19200 35000 9975 12000 ...
 $ Interest.Rate                 : Factor w/ 275 levels "10.00%","10.08%",..: 263 40 214 275 33 121 254 154 ...
 $ Loan.Length                   : Factor w/ 2 levels "36 months","60 months": 1 1 2 1 1 1 1 2 1 1 ...
 $ Loan.Purpose                  : Factor w/ 14 levels "car","credit_card",..: 3 3 3 3 2 10 3 2 2 2 ...
 $ Debt.To.Income.Ratio          : Factor w/ 1669 levels "0%","0.04%","0.17%",..: 390 1178 1000 346 657 775 ...
 $ State                         : Factor w/ 46 levels "AK","AL","AR",..: 37 39 5 16 28 7 19 18 5 5 ...
 $ Home.Ownership                : Factor w/ 5 levels "MORTGAGE","NONE",..: 1 1 1 1 5 4 5 1 5 5 ...
 $ Monthly.Income                : num  6542 4583 11500 3833 3195 ...
 $ FICO.Range                    : Factor w/ 38 levels "640-644","645-649",..: 20 16 11 12 12 7 17 14 10 16 ...
 $ Open.CREDIT.Lines             : int  14 12 14 10 11 17 10 12 9 8 ...
 $ Revolving.CREDIT.Balance      : int  14272 11140 21977 9346 14469 10391 15957 27874 7246 7612 ...
 $ Inquiries.in.the.Last.6.Months: int  2 1 1 0 0 2 0 0 1 0 ...
 $ Employment.Length             : Factor w/ 12 levels "< 1 year","1 year",..: 1 4 4 7 11 5 3 3 10 5 ...
```

I identified some missing value in the data set on the following factor variables *Home.Ownership*, *Employment.Length* and following numeric variables *Monthly.Income* , *Open.CREDIT.Lines, Revolving.CREDIT.Balance, Inquiries.in.the.Last.6.Months*

```
Amount.Requested Amount.Funded.By.Investors Interest.Rate     Loan.Length               Loan.Purpose
 Min.   : 1000    Min.   :   -0.01          12.12% : 122   36 months:1952   debt_consolidation:1307
 1st Qu.: 6000    1st Qu.: 6000.00          7.90%  : 119   60 months: 548   credit_card       : 444
 Median :10000    Median :10000.00          13.11% : 115                    other             : 201
 Mean   :12406    Mean   :12001.57          15.31% :  76                    home_improvement  : 152
 3rd Qu.:17000    3rd Qu.:16000.00          14.09% :  72                    major_purchase    : 101
 Max.   :35000    Max.   :35000.00          14.33% :  69                    small_business    :  87
                                            (Other):1927                    (Other)           : 208
 Debt.To.Income.Ratio     State       Home.Ownership Monthly.Income       FICO.Range
 0%     :   8         CA     : 433   MORTGAGE:1148   Min.   :   588.5   670-674: 171
 12.54% :   6         NY     : 255   NONE    :   1   1st Qu.:  3500.0   675-679: 166
 12.20% :   5         TX     : 174   OTHER   :   5   Median :  5000.0   680-684: 157
 12.85% :   5         FL     : 169   OWN     : 200   Mean   :  5688.9   695-699: 153
 14.22% :   5         IL     : 101   RENT    :1146   3rd Qu.:  6800.0   665-669: 145
 14.66% :   5         GA     :  98                   Max.   :102750.0   690-694: 140
 (Other):2466         (Other):1270                   NA's   :1          (Other):1568
 Open.CREDIT.Lines Revolving.CREDIT.Balance Inquiries.in.the.Last.6.Months Employment.Length
 Min.   : 2.00     Min.   :     0           Min.   :0.0000                 10+ years:653
 1st Qu.: 7.00     1st Qu.:  5586           1st Qu.:0.0000                 < 1 year :250
 Median : 9.00     Median : 10962           Median :0.0000                 2 years  :244
 Mean   :10.08     Mean   : 15245           Mean   :0.9063                 3 years  :235
 3rd Qu.:13.00     3rd Qu.: 18889           3rd Qu.:1.0000                 5 years  :202
 Max.   :38.00     Max.   :270800           Max.   :9.0000                 4 years  :192
 NA's   :2         NA's   :2                NA's   :2                      (Other)  :724
```

On average, at the same FICO score, the Interest rate is higher when the loan length is “60 months” (fig1a), (fig1), (fig2), (fig3).
Moreover the Interest Rate is

- slightly related to increased Amount Requested (fig6) and
- lower for some Loan Purpose factor like *renewable_energy*, *educational, car, home_improvement* (fig10)

I first fit a regression model relating Interest rate to FICOscore(avg between class), Debt-to-Income, Loan Length, open credit lines, (Revolving Credit Balance / Monthly Amount) ratio, HomeOwnership.

The residuals showed patterns of non-random variation. I attempted to explain those patterns by fitting models including potential confounder factors.

My final regression model was:

Interestrate = b0 + b1 FICOavg + b2 debt_income + b3 loanlenght60 months + b3 opencreditlines +
+ b4 emplenth + b5 inquiries + b6 monthlyincome + e

Where
Interestrate = Interest rate of the loan
FICOavg = average value between lower and higher class
debt_income = debt to income ratio
loanlenght60 months = loan length (factor = 60 Months)
opencreditlines = Open Credit Lines
emplenght = transformation from factor to numeric of Employment Length
inquiries = Inquiries in the last 6 months
monthlyincome = Monthly Income
b0 = intercept
bi(i=1,6) = coeficients

It is possible to observe a highly statistically significant (P < 2e-16) association between Interest Rate and both FICO range average and loanlenght60 months.

```
Call:
lm(formula = interestrate ~ FICOavg + debt_income + loanlength +
    opencreditlines + emplength + inquiries + monthlyincome,
    data = loansdata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.109364 -0.015898 -0.001684  0.013994  0.091297

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          7.121e-01  9.952e-03  71.561  < 2e-16 ***
FICOavg             -8.470e-04  1.365e-05 -62.036  < 2e-16 ***
debt_income          1.391e-02  6.992e-03   1.990   0.0467 *
loanlength60 months  4.267e-02  1.117e-03  38.207  < 2e-16 ***
opencreditlines     -1.905e-04  1.148e-04  -1.659   0.0972 .
emplength            2.998e-04  1.183e-04   2.533   0.0114 *
inquiries            2.936e-03  3.778e-04   7.770 1.15e-14 ***
monthlyincome        7.216e-07  1.225e-07   5.890 4.39e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0227 on 2413 degrees of freedom
  (79 observations deleted due to missingness)
Multiple R-squared:  0.7029,     Adjusted R-squared:  0.7021
F-statistic: 815.7 on 7 and 2413 DF,  p-value: < 2.2e-16
```

Conclusions:

The analysis suggests that there is a significant, negative association between Interest Rate and FICO range and a significant positive association between Interest Rate and Loan Length.

References:

[1] Wikipedia "Peer-to-peer lending". URL: http://en.wikipedia.org/wiki/Peer-to-peer_lending. Accessed on 15/11/2013.

[2] LendingClub "Better Rates". URL: https://www.lendingclub.com/public/about-us.action. Accessed on 15/11/2013.

[3] R Core Team (2012). "R: A language and environment for statistical computing." URL: http://www.R-project.org

[4] Makridakis, Wheelwright, McEGEE (1983). *Forecasting, Methods and Applications*, 2nd ed., Wiley

[5] Tutorials: Multiple Regression. URL: http://ww2.coastal.edu/kingw/statistics/R-tutorials/multregr.html

[6] R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed on 13/11/2013