Higher FICO score is associated with lower interest rate

Introduction:

Lending Club is a US peer to peer lending company and the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC). Lending Club allows loaning users to take up loans up to $35,000USD and allows lending users to "invest" in the loan by funding it; lending users will receive "return on investment" from the repayment of loans with interest. As of February 11, 2013, Lending Club has handled over 1 billion USD in loans (LC, Lending Club Statistics, 2013).

Lending Club's interest rates take into account credit risk and market conditions based on the equation: Lending Club Base Rate + Adjustment for Risk & Volatility (LC, Interest Rates and How We Set Them, 2013). "Adjustment for Risk & Volatility" is determined by the Lending Club on the risk of default. While Lending Club list "FICO score" (Wikipedia, FICO score, 2013), "Requested Loan Amount" and "Loan Maturity" as risk modifiers for formulation of "Adjustment for Risk & Volatility" (LC, Lending Club Statistics, 2013), one may guess there might be more to the formulation. Perhaps Lending Club also considers other variables such as employment history, credit history etc.

Thus, it will be an interesting problem to identify and quantify associations between the interest rate of the loan and the other variables in the data set. In particular, I will like to find out if 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?

Methods:

Data Collection

The data consisted of a sample of 2,500 peer-to-peer loans issued through the Lending Club. The data consisted of 14 columns: Amount.Requested, Amount.Funded.By.Investors, Interest.Rate, Loan.Length, Loan.Purpose, Debt.To.Income.Ratio , State, Home.Ownership, Monthly.Income, FICO.range, Open.CREDIT.Lines, Revolving.CREDIT.Balance, Inquiries.in.the.Last.6.Months and Employment.Length. The data was downloaded on February 9th, 2013 using the R programming language (R, 2013).

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. By observing these tables and plots, I identified transformations to perform on the raw data. Exploratory analysis was used to (1) identify missing values, (2) verify the quality of the data, and (3) determine the terms to be used in the regression model relating interest rate to FICO

rating. There were 7 missing values across variables like “monthly income”, “open credit lines”, “revolving credit balance” and “inquiries in last 6 months”. As they were all numeric variables, I imputed the values with the average of their respective variable. Interest rate and debt to income ratio was using “%” symbol, which was not identified by R to represent percentage. I had to drop “%” symbol from both columns and recoded the column as numeric. FICO.Range was dropped in favor of FICO Mean, as FICO Mean will be more suitable for graph plotting. Amount.Funded.By.Investors was dropped since it would not affect interest rate, as stated on lending club website (LC, What if my loan does not get fully funded?, 2013).

Statistical Modeling

From exploratory analysis, it was clear that simple linear regression would be a poor choice in helping me understand the relative importance of 14 different variables. I needed a stronger model than simple linear regression that could cater well to the multivariate data. I chose gradient boosting (scholarpedia, 2013), as it is a well known statistical model that can handle multivariate, non linear data with good accuracy. I used the R library GBM, which “implements extensions to Freund and Schapire’s AdaBoost algorithm and J. Friedman’s gradient boosting machine. Includes regression methods for least squares, absolute loss, logistic, Poisson, Cox proportional hazards partial likelihood, multinomial, t-distribution, AdaBoost exponential loss, Learning to Rank, and Huberized hinge loss.” (Ridgeway, 2013). Using the GBM library, I employed least squares regression using J. Friedman’s gradient boosting machine modeling (Friedman, 2001), as least squares regression can handle the both numeric and factor data formats for non integer variables.

GBM, like all models, has the possibility over fitting or under fitting the data. To mitigate the risk, I performed GBM with 5-fold cross validation (CMU, 2013). Using the results from 5-fold cross validation and GBM library function gbm.perf(), I was able to estimate the optimal number of boosting iteration for the GBM model. After performing GBM, I made use of GBM library function gbm.Summary() to obtain a table that lists the reduction of squared error attributable to each variable. This helped me determine the relative contribution of each variable in determining the interest rate (see results in table1 on next page).

Reproducibility

All analyses performed in this manuscript could be reproduced using the R markdown file ficoFinal.Rmd (RStudio, 2013) (available on request). To reproduce the exact results presented in this manuscript, the analysis must be performed on the same data set (available on request).

Results:

Table1: Variable contribution table

	Attribute	Contribution (%) in predicating interest rate
1	FICO Mean	61.270527
2	Loan Length	17.242917
3	Amount Requested	11.456389
4	State	3.5725721
5	Inquiries in the Last 6 Months	2.1737331
6	Open CREDIT Lines	2.1583824
7	Loan Purpose	0.5794753
8	Employment Length	0.5729936
9	Debt To Income Ratio	0.3803641
10	Monthly Income	0.2569005
11	Revolving CREDIT Balance	0.2309556
12	Home Ownership	0.1047900

From Table1 above, it was clear that the top three attributes – FICO Mean, Loan Length and Amount Requested contributed almost 90% in the prediction of interest rate and should be the key focus of my analysis.

Mean square error calculation (Wikipedia, Mean squared error, 2013)

To verify my results, I split my initial data set of 2,500 observations via random sampling into training (1,875 observations) and test data sets (625 observations). I then ran my GBM model through the test data sets and compared the predicted interest rate with the actual observed interest rate. Next, I calculated mean square error (MSE) using the formula:

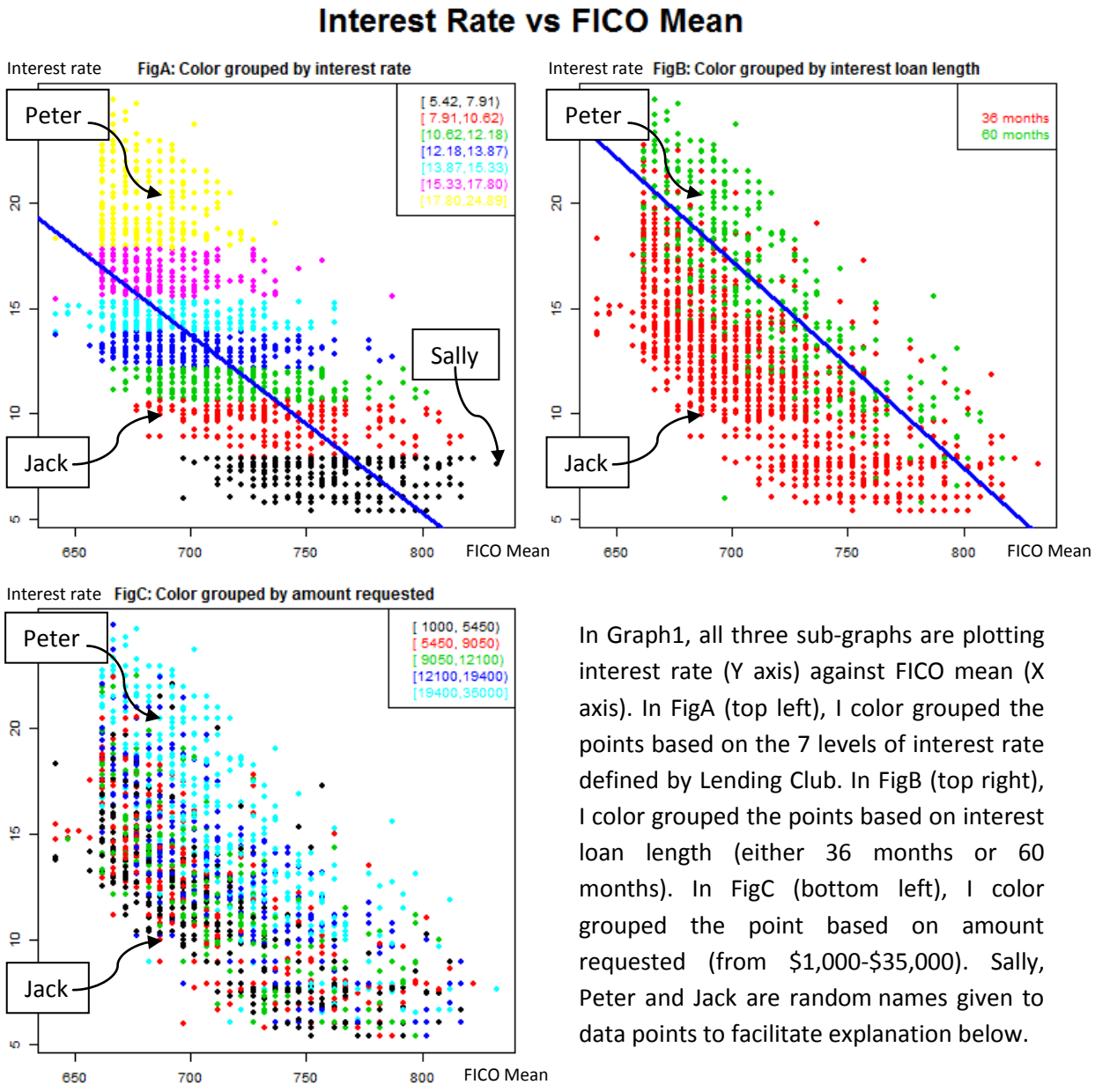
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

I obtained a root mean square error of around 1.88. Since the average interest rate of the data set was around 13.27, we are looking at an error margin of around 14% (1.88/13.27 * 100%). This was a fairly high error margin but I believe that the error rate will decrease if I had more training data. Furthermore, as the top 3 variables had a clear led over the other variables, an error margin of around 14% should not affect my final conclusion.

Confounders

Using ANOVA, I identified "Amount.Requested", "Loan.Length", "Loan.Purpose", "Debt.To.Income.Ratio", "Home.Ownership" and "Inquiries.in.the.Last.6.Months" as factors that are strongly associated with interest rate and fico mean. Going through a sanity check, however, it is likely that only "Debt.To.Income.Ratio" and "Inquiries.in.the.Last.6.Months" are logically and directly correlated with fico mean and might be confounders.

By plotting the top 3 attributes – FICO Mean, Loan Length and Amount Requested against interest rate, I arrive at the Graph1: Interest rate vs FICO Mean



In Graph1, all three sub-graphs are plotting interest rate (Y axis) against FICO mean (X axis). In FigA (top left), I color grouped the points based on the 7 levels of interest rate defined by Lending Club. In FigB (top right), I color grouped the points based on interest loan length (either 36 months or 60 months). In FigC (bottom left), I color grouped the point based on amount requested (from $1,000-$35,000). Sally, Peter and Jack are random names given to data points to facilitate explanation below.

From FigA, I saw that as expected, Sally who had a FICO mean score of more than 800 borrowed at a much lower interest rate than Peter and Jack, both who had a FICO mean score of less than 700. However, since Peter and Jack had the same FICO mean score, it was clear that FICO mean score was not the only determinant of interest rate. Looking at FigB, I could see that in general, users who applied for the longer loan length of 60 months (Peter) paid higher interest than 36 months loaner (Jack). Finally, from FigC, I saw that in general, users who applied for larger loan (Peter), paid higher interest than those with smaller loans (Jack).

The blue line on FigA represents a best fit linear regression line that relates interest rate with FICO mean. The blue line on FigB represents a best fit linear regression line that relates loan length, interest rate and FICO mean; hence you can see the line having a different gradient from FigA and the plot seemed separately into two sections by loan length. For FigC, I constructed another linear regression relating amount requested, loan length, interest rate and FICO mean. By looking at the P value from the summary statistics of this linear regression, I could deduce that interest rate is statistically significant in its relation to FICO mean and amount requested as P value is very small (<0.01) at 95% confidence interval. Also, for every point gained in FICO mean, the user will enjoy a decrease of interest rate by around 0.08 while every additional $1,000 requested by user increase the interest rate by around 0.36.

Conclusions:

My findings confirmed that the information provided by Lending Club on their website regarding the formulation of "Adjustment for Risk & Volatility" is probably valid (refer to introduction). Indeed, by referring to my Table1: variable contribution table above, I could see that a combination of "FICO Mean", "Requested Loan Amount" and "Loan Maturity" was strongly associated with interest rate value. What was more surprising, however, was how little the other variables actually mattered. One could argue that since "FICO Mean" is based on repayment timeliness, FICO implicitly accounted for variables like "inquiries in last 6 months", "open credit lines", "employment length", "debt to income ratio", "monthly income", "revolving credit balance"; hence their relative irrelevance to interest rate value.

On the other hand, I do still expect "loan purpose" to play a greater role in determining interest rates, since non-essential swimming pool loans might be at higher risk of default than essential loans like housing loan. Different "states" may also have different rules and regulations governing interest rates. Finally, while the current loan cap for Lending Club is limited to only $35,000, should the company decides to increase the loan cap limit, "Home Ownership" may become an important backing for loans and determinant for interest rate. Lending Club is still a rapidly growing company and its interest rate policies may still undergo numerous changes in the future. While "loan purpose", "state" and "Home Ownership" do not play any statistically significant role in determining Lending Club's current interest rate, we should watch out for any changes based on these 3 variables in any future works involving Lending Club. We should also involve a greater training data set in future works to ensure a greater accuracy for the trained models.

Bibliography

1. CMU. (2013). *Cross Validation*. Retrieved from http://www.cs.cmu.edu/~schneide/tut5/node42.html

2. Friedman, J. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. Retrieved from http://www-stat.stanford.edu/~jhf/ftp/trebst.pdf

3. LC. (2013). *Interest Rates and How We Set Them*. Retrieved February 11, 2013, from https://www.lendingclub.com/public/how-we-set-interest-rates.action

4. LC. (2013). *Lending Club Statistics*. Retrieved February 11, 2013, from https://www.lendingclub.com/info/statistics.action

5. LC. (2013). *What if my loan does not get fully funded?* Retrieved from http://www.lendingclub.com/kb/index.php?View=entry&EntryID=207

6. R. (2013). *R Project*. Retrieved from http://www.R-project.org

7. Ridgeway, G. (2013). *Package 'gbm'*. Retrieved from http://cran.r-project.org/web/packages/gbm/gbm.pdf

8. RStudio. (2013). *Using R Markdown with RStudio*. Retrieved from http://www.rstudio.com/ide/docs/authoring/using_markdown

9. scholarpedia. (2013). *Boosting*. Retrieved from http://www.scholarpedia.org/article/Ensemble_learning#Boosting

10. Wikipedia. (2013). *FICO score*. Retrieved February 11, 2013, from http://en.wikipedia.org/wiki/Credit_score_in_the_United_States#FICO_score

11. Wikipedia. (2013). *Mean squared error*. Retrieved from http://en.wikipedia.org/wiki/Mean_squared_error