The FICO score is not the only determinant of the credit interest rate in the Lending Club

Introduction:

The Lending Club is the world's largest peer-to-peer lending platform, an online financial community that brings together borrowers and investors so that both can benefit [1]. Generally speaking, the interest rate at which the money is lent is determined by the risk-free interest rate plus an adjustment for risk and volatility. In the United States, this adjustment is usually calculated with, among other variables, the FICO score, which is a measure of creditworthiness provided by a public company [2].

A better understanding on how the Lending Club actually calculates the interest rate can be very helpful both for the borrowers and for the investors, increasing the transparency in the market. This kind of information is especially important for borrower, since it may allow them to estimate and anticipate the cost of a particular credit. At the same time, it may allow them to strategically influence some variables that are taken into account when setting the final interest rate.

Here we perform an analysis to determine which variables are relevant when determine the interest rate of a loan and to quantify these relations. We used explanatory analysis and multiple regression techniques to show that there are several variables, apart from the FICO score, that affect the final interest rate. Moreover, not all the variables affect linearly, some of them turned out to be more relevant for the lower levels of creditworthiness.

Methods:

Data Collection

We have used a sample of 2,500 peer-to-peer loans issued through the Lending Club for our analysis. The data was downloaded on February 4th, 2013 form https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv using the R programming language [3]. The code book for the variables 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 sample data. The analysis was used to verify the quality of the data, including the identification of missing values, and to determine the terms used in the regression model.

As to the quality of the data, a total of 88 values were missing, although finally only 2 observations were discarded due to missigness. The other missing values were placed in factorized data so we transformed it into another category.

Statistical Modeling

To relate the interest rate of the loans to the characteristics of the borrower and the credit we used a standard multivariate regression technique [4]. Our previous exploratory analysis and the prior knowledge of the market were key in determining the model selection. Some non-linearity in the relations among the interest rate and the other variable were allowed by adding the square variable to the regression. The regression was estimated using ordinary least square. [5]

Reproducibility

The analysis performed in this manuscript is reproducible in the R markdown file [6], although the file is not provided due to security concerns.

Results:

The loan data used in this analysis contains information on the specific credit and information about the borrower. The variables that have to do with the credit were the interest rate (IR), the amount requested (AR, measured in dollars), the amount finally funded by investors (AFI, measured in dollars), the length of time of the loan (LL, that can be either 36 or 60 months) and the purpose of the loan (PL). The variables that have to do with the borrower were the state of residence (SR), the percentage of consumers' gross income that goes toward paying debts (DIR), a variable indicating whether the applicant owns, rents, or has a mortgage on their home (HO), the applicant's monthly income (MI, in dollars), a range indicating the applicant FICO score (FICO), the number of open lines of credit the applicant had (OCL), the total amount outstanding all lines of credit (RCB, in dollars), and the inquiries about their creditworthiness that the applicant had authorized in the six months previous to the loan issuing (I6M), and the employment length (EL, measured as a range from less than one year to 10 or more years).

In order to obtain a direct relation between the interest rate and the FICO score and to increase the degrees of freedom in the regression, the initial FICO range was transformed into a numerical variable using the average value between the top and the bottom range. Subsequent analyses focus on this transformed debt variable.

We first fitted a regression model relating linearly interest rate to FICO and other variables that may potentially have an effect. However, the residuals showed patterns of non-random variation and, therefore, we introduced the quadratic expression of some variables. We also discarded all those variables that were not statistically significant. Our final model was:

$$IR = a + b_1(FICO) + b_2(FICO)^{\wedge 2} + b_3(AR) + b_4(LL) + b_5(S) + b_6 (HO) + b_7 (OCL) + b_8(OCL)^{\wedge 2} + b_9 (I6M) + e$$

where a is an intercept, b_i represents the effect that the different explanatory variable have on the interest rate, and e represents the error term, including all sources of unmeasured and unmodeled random variation in the interest rate. The final regression appears to remove most of the non-random patterns of variation in the residuals.

As expected, we found a clear and strong relationship between the interest rate and the borrower's FICO. But this relation appears to be better fitted in a quadratic relation, where the lineal parameter is negative (-0.85 with P<0.001) and the quadratic parameter is positive (0.00053 with P<0.001). Since the FICO score of the sample expands between 642 and 832, it implies that, all other thing equal, an increase in one point in the FICO score reduces more the interest rate to those applicants with a lower FICO score.

The amount requested turns out to be also an important driver of the interest rate (0.00016 with P<0.001). This means that, all other variable equal, requesting 1,000 extra dollars is associated with a 0.16 percentage points (pp) interest rate increase. Analogously, demanding a

60 months loan instead of a 36 months implies that the interest rate increases in 3.3 pp (with P<0.001)

The number of open credit lines also affects the interest rate in a non-linear manner. In particular, the lineal parameter is negative (-0.43 with P<0.001) and the quadratic parameter is positive (0.016 with P<0.001). Therefore, all other things equal, the minimum interest rate is charged to those borrowers with around 14 open lines of credit, while the penalization for those that only has two lines (the minimum in the sample) is more or less the same than of those with 25 lines.

Another relevant variable is the number of inquiries about the creditworthiness of the applicant over the previous six months. In particular, one extra inquire implies, *ceteris paribus*¸ 0.34 extra pp (with P<0.001). With regard of the applicant's home status, it turned out that it is only significant variable for those who rent their home (0.026 with P = 0.0021).

Finally, the State of residence of the loan applicant has also an impact on the interest rate, being the states of Iowa and Indiana the cheapest and New Mexico and Wyoming the most expensive. However, this relation should be taken with caution, since there could be some confounding variables correlating with the interest rate and the characteristics of the average borrower from different States. This could imply a bias in the coefficients and their statistical relevance.

It is also worth noting that there were some variables that were *a priori* candidates that resulted in no statistical significance and did not enter the final model specification. In particular, we highlight variables such as the purpose of the loan, the monthly income, the total amount outstanding all lines of credit, the debt to income ratio, the total amount outstanding all lines of credit and the employment length. One explanation may be that some of this information is already taken into account when constructing the FICO variable and are omitted afterword to avoid double counting.

Conclusions:

Our analysis suggests that there is a significant relation between the interest rate charged for the loans and the FICO scores of the applicant. Moreover, this relation appears to be better fitted with a non-lineal model that suggests that increases in the FICO score have a greater impact for those with a bad credit rating than for those with a good score.

We also observed that there are several other variables that affect the interest rate, namely the total amount requested, the length of the loan, the number of open credit lines, the number of inquiries about the creditworthiness of the applicant over the previous six months and the State of residence.

This analysis may constitute a good first step in order to shed some light on how the interest rate of a credit is calculated. However, it has been performed using data only from the Lending Club, which can work quite differently from other financial institutions. Therefore, the conclusions outside this lending platform may not be valid and should be extrapolated very cautiously.

Moreover, the data used corresponds to people who had actually been qualified for the loan and therefore all applicants have already passed an initial credit screening criteria. In order to get a better understanding on how this platform works, it would be also interesting to have data from rejected applicants.

References:

1. The Lending Club Page. URL: https://www.lendingclub.com/home.action. Accessed 02/12/2013.
2. Wikipedia “FICO score” Page. URL: http://en.wikipedia.org/wiki/FICO_score#FICO_score. Accessed 02/12/2013.
3. R Core Team (2012). ”R: A language and environment for statistical computing.” URL: http://www.R-project.org. Accessed 02/12/2013.
4. Wikipedia “Linear regression” Page. URL: http://en.wikipedia.org/wiki/Linear_regression. Accessed 02/12/2013.
5. Wikipedia “Ordinary least squares” Page. URL: http://en.wikipedia.org/wiki/Ordinary_least_squares. Accessed 02/12/2013.
6. R Markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 02/12/2013.