

Title: Dependency of various applicant attributes with loan interest rate

Introduction:

In this analysis of LendingClub [1] loans dataset, we identify associations between interest rates and other characteristics of persons asking for a loan such as loan length or requested amount. We evaluate correlations between attributes, we plot several graphs displaying relations between parameters and we fit multiple linear models to quantify attribute dependency.

Such analysis can be useful in interest rates predictions or it can be used to ease interest rate estimation from characteristics of person asking for the loan.

Methods:

Data collection and preprocessing

For our analysis we used a dataset of 2500 peer-to-peer loans issued through the LendingClub [1]. The dataset consists of 14 attributes such as interest rate (numeric), loan length (factor) or open credit lines (integer). We transformed several attributes to different types using R language [2]. For example we transformed vector of interest rates from character vector to numeric to improve data manipulation possibilities.

We defined ordering of levels for factor attributes (employment length). We explored missing data and we imputed them using nearest-neighbor (kNN) [3] method.

Exploratory analysis

Exploratory analysis was performed by examining means, quartiles, tables and histograms for different attributes. By studying individual attributes we were able to verify quality of the data and find the most important attributes for further analysis. We created correlation matrix, we performed hierarchical clustering and we plotted different combinations of attributes to determine relations between attributes. We selected attributes with highest correlation with interest rate for further analysis. We created plots of interest rate and the most correlated attributes. We used linear model fitting [4] to quantify dependency between attributes.

Statistical Modeling

We performed linear model fitting on interest rate and FICO score for two groups of loan length attributes. Linear model attributes selection was based on correlations between interest rate and other dataset attributes.

Results:

The loans data used in this analysis contains of characteristics of person asking for a loan as well as information about the loan itself: amount requested in loan application

(Amount.Requested), amount loaned to the individual (Amount.Funded.By.Investors), interest rate (Interest.rate), length of the loan (Loan.length), loan purpose (Loan.Purpose), percentage of individual income, that goes toward paying debts (Debt.to.Income.Ratio), abbreviation of U.S. state of residence of the loan applicant (State), whether applicant own/rent/... home (Home.ownership), applicants monthly income (Monthly.income), range indicating applicants FICO score (FICO.range), number of open lines of credit at the time of application (Open.CREDIT.Lines), total amount outstanding all lines of credit (Revolving.CREDIT.Balance), number of authorized queries about applicant creditworthiness for last 6 months (Inquiries.in.the.Last.6.Months), length of time applicant is employed at current job (Employment.Length).

We identified several missing values (87) and multiple outliers (for example negative Amount.Funded.By.Investors). We removed outliers by sampling and leaving out these rows. We resolved missing values by inputting them using nearest-neighbor (kNN) [3] method

We determined dependency between attributes using correlation matrix, hierarchical clustering and plotting various parameter combinations. We ordered correlations with interest rate by decreasing absolute value and we found several attributes highly correlated with interest rate (FICO.range, Loan.Length, Amount.Funded.By.Investors and Amount.Requested).

We displayed dependency of interest rate and FICO score using boxplot, where inverse dependency is clearly visible. We plotted interest rate and FICO score using scatterplot where data points were colored by different loan lengths (2 groups). We fitted a linear model to determine relation of interest rate with FICO score and loan length. The final formula for the linear model was:

$$\text{Interest.Rate} = b_0 + b_1 (\text{Loan.Length}_i = "60 \text{ months}") + b_2 \text{Loan.Length}_i + b_3 \text{Loan.Length}_i * (\text{Loan.Length}_i = "60 \text{ months}") + e_i$$

Where:

- b_0 – interest rate at FICO score equals to 0 for loan length equal to 36 months
- $b_0 + b_1$ – interest rate at FICO score equals to 0 for loan length equal to 60 months
- b_2 – change in interest rate (loan length equal to 36 months) in one FICO score point
- $b_2 + b_3$ – change in interest rate (loan length equal to 60 months) in one FICO score point
- e_i – all sources of unmeasured random variation in interest rate

We observed statistically significant association between interest rate, FICO score and loan length ($P < 2.2e-16$).

We plotted interest rate and FICO score using scatterplot where data points were colored by increasing amount funded by investors. The color shifts from orange to blue by increasing amount funded by investors. The clear separation of blue and orange points indicates similar dependency of interest rate, FICO score and amount funded by investor as in previous example.

Conclusions:

Our analysis suggests that there is a significant association between interest rate, FICO score and loan length. The graph displaying relation between interest rate, FICO score and amount funded by investors suggest similar association. Using method such as correlation matrix, we observed that other parameters correlate with interest rate to some point. In performed analysis we used only limited data sample, but it suggests interesting relations between applicant attributes and loan interest rate. This preliminary analysis may serve as first step in further association and pattern exploration.

References:

1. LendingClub Page. URL: <https://www.lendingclub.com/home.action>. Accessed 11/17/2013
2. The R Project for statistical computing Page. URL: <http://www.r-project.org/>. Accessed 11/17/2013
3. Documentation for method to impute missing data. URL: <http://svitsrv25.epfl.ch/R-doc/library/impute/html/impute.knn.html> Accessed 11/17/2013
4. Linear Model Fitting Documentation. URL: <http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lm.html> Accessed 11/17/2013