Lending Club Data Analysis and Default Rate Prediction

Proposed by Qing Zhao

Lending Club (LC) is the world's largest online marketplace connecting borrowers and investors. It is transforming the banking system to make credit more affordable and investing more rewarding. Lending Club operates at a lower cost than traditional bank lending programs and pass the savings on to borrowers in the form of lower rates and to investors in the form of solid risk-adjusted returns.

Details of how it works can be found here.

OBJECTIVE

In this project, you are expected to play with the data provided by LC, conduct a set of exploratory analysis and try to apply various machine learning techniques to predict borrower's default rate.

DATA

All data regarding this project can be accessed here. So far the data is available till 2016 Q2 and can be dated back to 2007. The data consists in 4 files updated every quarter on the same day as the quarterly results of the company are released. They contain information on almost all the loans issued by LC. The only loans missing from these files are the few loans where LC was not authorized to release publicly the details of the transactions.

The information available for each loan consists of all the details of the loans at the time of their issuance as well as more information relative to the latest status of loan such as how much principal has been paid so far, how much interest, if the loan was fully paid or defaulted, or if the borrower is late on payments etc.

LOAN DATA

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. You need to register an LC account and sign in to download the full version of the files.

- DECLINED LOAN DATA

  These files contain the list and details of all loan applications that did not meet Lending Club's credit underwriting policy.

There are 115 data attributes in total, however, not all of them can contribute to data analysis or can be used as features for machine learning techniques. It is your responsibility to identify which features to use, or even better, create your own features based on the given attributes.

MILESTONES

This part serves as a suggested project planning. Some parts can be later adjusted through discussions with TF.

1. **Project Selection**

   Form teams of 2 or 3 and select a project from the provided list.

2. **Data Cleaning and Exploratory Analysis**

   The data provided in this project is relatively clean and formatted, however, you may need to deal with some null values or transform some categorical variables. You are also encourage to seek for external data sources.

   Based on the preprocessed data, you are expected to conduct a series of exploratory data analysis to get a taste of how the data looks like or how it is distributed. **In this part you are required to perform various data visualisation to show what you have found.**

3. **Feature Engineering**

   Since there more over 100 data fields (115) provided, it is not necessary, even not suggested to use all of them. It is your job to perform feature selection and discover the features that are indicative of someone paying or defaulting on their loan.

4. **Predict Default Rate**

   Here comes the most exciting part! Based on the previous data exploration and feature engineering, you can predict whether someone will default based on some classification algorithms or predict someone's default possibility using various regressions so as to avoid loans that are predicted to default.

   Possible algorithms include but not limited to: Logistic Regression, SVM, Naive Bayes, Linear Regression, Random Forest, Ada Boost, etc.

5. **Model Comparison and Evaluation**

   Here you are expected to come up with a set of evaluation criteria to compare different model's performance and discuss their "Bias vs Variance". It is important to try to explain why the models give such results and what are their pros & cons under certain constraints.

6. **Final Report**

As the last step, please write a final report in iPython notebook to include all your codes, graphs and writeup.

REFERENCE

Here are some works done previously on this topic, feel free to take a look and grab things you think could be helpful to your work. You are also encouraged to research more on related work online. However, you **MUST** come up with your own ideas and mark all your references at the end of your final report.

- https://www.lendingclub.com/info/statistics.action
- http://www.lendingmemo.com/lending-club-prosper-default-rates/
- http://res.cloudinary.com/general-assembly-profiles/image/upload/v1416535475/uwumoooppttsmpgu1goo.pdf
- http://nbviewer.jupyter.org/gist/odubno/0b767a47f75adb382246