Predicting Loan Outcomes using Machine Learning

Abhishek Malali

August 25, 2016

Abstract

Financial companies who have interests in personal loans face risk from customers who default. With the advent of predictive analytics, the risk to the company can be reduced by being able to predict the outcome of these loans. The personal data collected by the companies during loan applications is a goldmine of information. Mining the data to create powerful features which are able to predict the loan outcome, can prevent write offs for lenders. This increases profitability and can help maintain a healthy lending market. With the data, the interest rate and loan grade prediction can be useful to businesses as well. This allows for faster processing of loans for the customers and lenders.

Milestones

Problem Statement

Multiclass classification problem for predicting loan status from a rich dataset. The problem can be reduced to a binary classification problem to build a loan approval pre-check system for potential customers.

Literature Survey

Loan default trends have been long studied from a socio-economic stand point. Most economics papers believe in empirical modelling of these complex systems in order to be able to predict the loan default rate for a particular individual. The use of machine learning for such tasks is a trend which we are observing now. Some of the papers to understand the past and present perspective of loan defaults are -

- Serrano-Cinca C, Gutirrez-Nieto B, Lopez Palacios L (2015), Determinants of Default in P2P Lending. PLoS ONE 10(10)
  URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0139427

Data

The data is available on Kaggle Datasets. The data can be downloaded from here. In addition to the loan dataset, Kaggle provides a data dictionary explaining the column headers in much more detail allowing much better understanding of the problem at hand.

Exploratory data analysis

Some of the examples of exploratory data analysis are:

- Visualize the distribution of loan amounts to understand lending trends over time.

- Extensive data driven study of the reasons of people taking the loans and ascertaining reasons for the default.
- Study and visualize the demographic nature of loan defaults.
- Computing and validating correlations between load status and indicators like FICO score, credit score and other personal history in order to model interactions.
- Visualizing the correlation between interest rates and loan grades. Correlating the interest rate with demographic data will be an interesting task.

Proposal

The interesting problems an investigator might face with this dataset include

- Practical and usable cross-validation strategy.
- Choosing evaluation metric for the dataset and the problem being solved.
- Handling missing data.
- Building features from the rich text data available in loan descriptions.
- Example solutions include:
  - Predicting loan status from loan grades.
  - Using logistic regression to predict loan status(In case problem is being reduced to a binary classification problem).
  - Using LDA to parse the textual data and categorize loans into certain categories from available information.
  - Building a tree or neural network based solution and generating the most important features.