Topics in Quantitative Finance - ML: Project Proposal

Michal Topinka

March 27, 2017

Loan Default Prediction

In this project, I will use data from a three-year-old competition on *kaggle.com*. (`https://www.kaggle.com/c/loan-default-prediction`) The dataset contains a set of standardized, de-trended and anonymized financial transactions with over two hundred thousand observations and 778 features. For each observation, it was recorded whether a default was triggered. In case of a default, the loss was measured. The dependent variable (*loss*) represents the percentage of the loan that was not repaid by the debtor and therefore takes values 0 to 100. The rest of the features is labeled from *f1* to *f778*, so we don't know what the particular features represent. The task of the competition was to predict whether a loan will default, as well as the loss incurred if it does.

In my project, I will attempt to solve the same problem and will see how close can I get to the result of the competition's winner using the methods learned in this course. The recommended approach here is to tackle this problem using a two-step process, where the first step is to use some classification methods to determine whether a person will default and then using regression analysis to predict the size of the loss (percentage). I will also try to identify the most important features and perhaps even try to create some new features from those that are already given, as implemented by some of the competition's winners. For the classification, I will try to use methods such as logistic regression, support vector machine, decision trees or k-nearest neighbors. To determine the size of the loss I will start with a classic linear regression, where the biggest problem will be to select the appropriate features to use in the regressions.

Since this is a closed competition, there are already winners whose code is published on the website. I keep this in mind, but I still believe that there is a lot to be done with such a large pre-processed dataset and that there are many different how to address this problem, so I will try to implement some of them.

The full dataset, as well as the codes of the winners and additional information, is available on *kaggle.com*, using the link in the first paragraph.