Group member: Xiaofeng Cao

Preliminary title: Predicting how capable each applicant is of repaying a loan

Credit risk is one of the most critical risks that a bank faces when it issues a loan to a borrower. It is extremely important for loan issuers to be able to identify risks as well as evaluate risks. There are plenty of risk models available in the market. However, most of these risk models are designed to evaluate creditworthiness of businesses (nonfinancial firms, banks, insurance companies etc.) Measuring creditworthiness of individuals replies more on qualitative analysis since information on individuals is not easy to achieve. Internet booming benefits us to gather more useful data on business purpose. In this research, we will apply data mining techniques on a set of loan applicants’ data to predict applicant’s capability to repay a loan. This study will significantly improve efficiency on issuing a loan to customers. The prediction result of the model will serve as a guide to loan managers.

Predicting applicant’s capability to repay a loan is the primary issue to solve. This is a typical classification problem. The `TARGET` feature in the application\_train.csv is a binary feature to predict. The potential problems in the research would be class imbalance and features selection since probability of default on a loan is relatively low comparing to flip a coin. I will probably apply decision tree, random forest, logistic regression and more to the data. The primary evaluating method would be confusion matrix (f1, precision, recall and accuracy) and roc curve. Applying cross validation is critical to success a machine learning model. I will look at the accuracy curve to avoid overfitting.

I will use the Home Credit Default Risk dataset from Home Credit Group. This is a closed competition on Kaggle. There are 7 different sources of data. The `TARGET` feature is in application \_train.csv. The training data has 307511 observations (each one a separate loan) and 122 features (variables) including the TARGET (the label we want to predict). I can imagine myself spending a lot of time on joining these files and on selecting important features from these files. My primary coding tool will be Python (pandas, sklearn, and matplotlib). The 7 different data are listed below.

* application\_train/application\_test: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK\_ID\_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.
* bureau: data concerning client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.
* bureau\_balance: monthly data about the previous credits in bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.
* previous\_application: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK\_ID\_PREV.
* POS\_CASH\_BALANCE: monthly data about previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.
* credit\_card\_balance: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.
* installments\_payment: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.