Loan Default Prediction

Michal Topinka

Peking HSBC Business School

1602010585@sz.pku.edu.cn

April 20, 2017

Data Source

- The data come from a three-year old competition on *Kaggle.com* hosted by the Imperial College London
- The goal was to determine whether a loan will default as well as to estimate the loss incurred if it does
- More specifically, they had to predict the loss for each row in the test set (next slide)
- The competition is closed now, winners were announced and the three people who were ITM had to expose their code

Data Description

- The dataset correspond to a set of financial transactions associated with individuals
- The data have been standardized*, de-trended, and anonymized. There are 769 features labeled f_1 to f_{769}
- 2 datasets:
  - Training set - 105471 samples, including dependent variable 'loss'
  - Test set - 210944 samples, **without the variable 'loss'**
- The 'loss' variable ranges from 0 to 100
  - 0 = no default
  - 70 = only 30% of the loan was reimbursed
  - 100 = 100% of the loan was not repaid

Project

- Since the data for the 'loss' variable for the test set was not published, there is no way for me to check the accuracy when using the test set $\rightarrow$ **I will only work with the training set**
- **Project outline**
  - Data preprocessing (checking for duplicate rows, duplicated and constant columns, imputing NA values, etc.)
    - $\rightarrow$ removed 40 columns, 0 rows
  - Feature Selection
  - Feature Extraction
  - Classification (Linear regression, SVM)
  - Regression analysis to predict the size of the loss

Methods (already implemented)

- Logistic Regression
- SVM (computational problems)
- Random Forrest Classifier