CS 6220 Summer '18 Project Proposal : Home Credit Default Risk

Chia Yi Liaw
liaw.c@husky.neu.edu

Mounica Subramani
subramani.m@husky.neu.edu

Sharyu Deshmukh
deshmukh.s@husky.neu.edu

Somya Bhargava
bhargava.so@husky.neu.edu

Problem Description

Home Credit is a finance provider that focuses on serving the unbanked population. Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders. So, Home Credit tries to broaden financial inclusions for unbanked population. The Home Credit Default Risk challenge is a standard supervised machine learning task where the goal is to use historical loan application data to predict their clients' repayment abilities based on datasets provided.

Data Exploration

The dataset consists of 1 main training data file (with the labels included), 1 main testing data file, and 6 additional data files as described below

- *application training/testing.csv*: contains the loan applicants which with the labels included a binary 0- if the loan was repaid- and a 1 if the loan was not repaid
- *bureau.csv*: application data from other credit institution with historical record
- *bureau_ balance.csv*: monthly balances of previous credits in Credit Bureau
- *previous_ applications.csv* : previous loans information at Home Credit by same clients
- *credit_ card_ balance.csv*: monthly balance snapshots of previous credit cards that the applicant has with Home Credit.
- *POS_ CASH_ balance.csv*: cash loans that the applicant had with Home Credit
- *installments_ payments.csv*: repayment history for the previously disbursed credits in Home Credit

Proposed Plan of Approach

- Data Preprocessing: Missing values will be examined and handled with imputation or deletion. The unbalanced distribution in target variable will be handled by applying over-sampling method. correlations between the features and the target will be examined to understand the data. Feature engineering will be performed by encoding the categorical variables. Feature selection will be done to analyze important features using techniques such as Random Forest, Gradient Boosting and PCA
- Modeling : Multiple machine learning models shall be applied on the dataset, such as Random Forest, SVM, XGBoost and Logistic Regression to achieve the in-depth data analysis, classification and prediction on client repayment abilities. Cross validation will be performed which includes setting hyperparameters and model training.

Desired Outcomes

- Evaluation : Models will be evaluated and compared using the accuracy score and Receiver Operating Characteristic Area Under the Curve. In addition to that, we plan to interpret the influential predictor.
- Visualizations : We plan to utilize Tableau and Python to deliver data visualization of model evaluation and prediction results.

References

- Home Credit Default Risk
  https://www.kaggle.com/c/home-credit-default-risk
- Predicting Customer Debt Default
  https://nycdatascience.com/blog/student-works/kaggle-predict-consumer-credit-default/