Credit Risk Modeling for Home Credit Group

Pushkar Sathe and Johann Thiel
Github Page

Introduction

Home Credit Group (HCG) is a company that specializes in lending to individuals with little to no credit history. Due to the nature of these loans, HCG is interested in determining the probability that an individual will default on a loan. Data was obtained from a Kaggle challenge[1] from HCG.

Problem and Metrics

- Construct a model that can find individuals who are likely to default given sparse data obtained by combining various internal and external sources of information.
- The model should be 'stable' in the long run and should not experience a severe degradation over time even at the cost of some short-term performance.
- Since defaults are rare events, we will compute ROC-AUC values/curves to determine model performance. Weekly gini score, $gini_w = 2AUC - 1$
- The stability score is a custom metric from HCG designed to penalize high-variance models that degrade over time. Stability metric = $mean(gini_w) + 88.0min(0,a) - 0.5std(r)$

Data and Cleanup

- The data set is large, ~26GB, with ~5.7 billion entries where ~41% of all entries are NaN, and spans 32 files from various sources with 465 different features that are a mix of numerical, date, categorical, masked values (to protect personal identification information).
- One approach taken was to concentrate on numerical features, allowing us to severely reduce the size of the data set (at the cost of discarding a lot of information).
- A second approach was to convert all of the features into categorical values, keeping all NaN values as a category to avoid discarding too much information.

Models

- Logistic Regression
- Decision Tree Classifier
- AdaBoost using decision tree classifiers
- Fully Connected Neural Network (FCN)

Results - validation data

- Model risk results: (ROC-AUC): Logistic regression: 0.56, AdaBoost: 0.78, Neural network: 0.82
- Stability results: AdaBoost: 0.5314, Neural network: 0.5698

Summary and Future Work

Two of our models were able to outperform the baseline model provided by HCG in the competition. Data munging is the most challenging step in this project. It can be refined further to improve the performance of all models. Furthermore, hyperparameter tuning can also generate gains in the stability metric and the ability to predict defaults.

[1] https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability/data