Home Credit Risk Stability II

Team members: Brady Ali Medina, Sze Hong Kwong

Github: https://github.com/kwongszehongumdmath/erdos-homecreditmodelstability

Goal:

- For each credit application, predict its likelihood of default.
- Maintaining stability: the model performance should not fluctuate over time.

Stakeholders: Consumer finance providers (e.g. Home Credit Group), individual loan applicants (particularly whose with little or no credit history)

KPIs: AUC, gini metric, and stability metric.

Data: Kaggle competition *Home Credit - Credit Risk Model Stability* (https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability/)

Data processing:

- The dataset has ~1.5 million rows, but with only ~3% of default cases
- It consists of internal and external tables
- Tables with depths>1 are provided with number group labels for aggregation
- Most columns are ready to be used for data analysis
- Imputations: filling in by zeros or medians; or add a missing indicator
- Features engineered:
  1) a collection of weighted dpd figures: higher weights are put on data that reflect long term behaviors.
  2) risk assessment from credit bureau, turned into a mean probability of default

Exploratory data analysis:

- Features are only weakly correlated with the target
- The weighted dpd figures improve correlation with target
- The generated mean risk assessment has correlation ~0.25, by far the highest, but only a small population has that assessment
- Summary statistics of basic static info (income, debts, etc.) are almost identical for default and non-default population. They are more distinguishable when stratified by income type.

Models:

- Our baseline model: XGBRF Classifier. It has AUC score 0.5, stability metric 0.0.
- We choose the following models: (i) XGBoost, (ii) MLP Classifier and (iii) LightLGB

Results:

- LightGBM offers the best stability across all datasets with consistent scores.
- All models outperformed the baseline

metric\model	**XGBRF Classifier (baseline)**	**XGBoost**	**MLP Classfier**	**LightGBM**
Train	-0.000	0.236	0.577	0.562
Validation	-0.000	0.209	0.109	0.534
Test	-0.000	0.165	0.151	0.521

Future Iterations:

- We are going to add the best features of our other models to the Light GBM.
- We are going to add external data such as the M2 Money Stock(measure of the total money supply in circulation).