Applied Machine Learning
Final Project:
Home Credit Default Risk

Group 16
Aravind Reddy Sheru, Sai Charan Chintala, Seongbo Sim and Yun Joo An

Indiana University

November 2021

Contents

Team Profile

Aravind Reddy Sheru
asheru@iu.edu

Sai Charan Chintala
sachin@iu.edu

Seongbo Sim
simseo@iu.edu

Yun Joo An
yunjooan@iu.edu

Four P's

- Past
  - We are making the HCDR Project, which predict whether borrowers are defaulters or not based on various financial and nonfinancial data.
- Present
  - In this phase, we collected the data, and did EDA. Also, we built a baseline model using logistic regression and tried to balance the data by adjusting the number of samples of non-defaulters.
  - The baseline model gave a quite high accuracy, but relatively low AUC. Balancing data improved AUC at the cost of accuracy.
  - We have learned that we lost informations of samples which were excluded in rebalancing, so we had to find better way to improve AUC while keeping the samples.

Cont'd

- Planned
  - In phase 2, we will introduce other candidate models, including "Decision Making Trees", "Random Forest", and "SVMs".
  - Also, we are planning to make our inputdata set more proper by feature engineering and feature importance analysis.
  - For candidate models, we will adjust hyperparameters to improve AUC and other metrics with higher accuracy.
  - Also, we are planning to adjust the details of the models, by Additional Feature Engineering.
  - Finally, we will ensemble our models to get a better results.
- Problems
  - We may need some prior knowledge about the data, for example credit data.
  - The knowledge may help us to evaluate the process of feature engineering, and feature importance analysis.

Final Project: Project Description

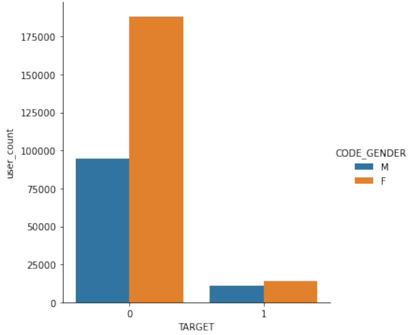
- The object of HCDR project is to predict the repayment abilities of financially under-served population.
  - The well-established prediction is necessary to both of Home Credit and borrowers.
  - "Lend money to whom can pay back" & "Give a chance to build credit".
- We use versatile data, e.g. previous credits information, type of credit, remaining days for previous credit, payments, previous application details, etc.
  - We utilize both of numerical and categorical data to increase the quality of prediction.
  - By EDA, we build proper dataset for machine learning models.

Cont'd

- To find the best model, we train and evaluate several candidate models.
  - Our candidate models are "Logistic Regression", "Decision Making Trees", "Random Forest", and "SVMs".
  - We use different evaluation metrics to have a concrete evaluation of candidate models including "Accuracy", "F1 Score", "AUC", and "KS-Score".
- Once the winning model is selected, we expect the model gives satisfactory prediction on the new test data.

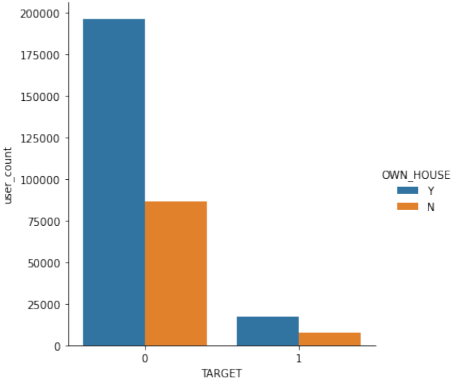
Final Project: EDA

- We do EDA, and check the following attributes of the data
  - "Test Description", "Dataset size", "Summary Statistics", "Correlation analysis", "Checking missing values", etc.
- Some interesting EDAs
  - Defaulters and Non-defaulters based on gender



Cont'd

- Defaulters and Non-defaulters with regard to home ownership



Final Project: Modeling Pipelines

- The specific object is predict whether the borrower is a defaulter or not.
- In Phase 1, we build a baseline model of "Logistic Regression".
    1. Split the dataset into data for training and that for testing.
    2. Prepare the input dataset (Scaling the data and converting missing values).
    3. Conduct "Logistic Regression."
    4. Evaluate the baseline model with "Accuracy" and "AUC" metrics.
    5. Do steps 1~4 with 50,000, 75,000 randomly selected non-defaulters in training data.
- Note that the number of non-defaulters is about 10(7) times greater than that of defaulters in training data (test data).
- The detailed setup is documented in the report, and omitted here.

Final Project: Results

- Results for baseline model

Model	Cross fold train accuracy	Test accuracy	AUC
Baseline (Logistic Regression)	91.9	91.9	0.502
Baseline with 50k non-defaulters	71.4	71.4	0.622
Baseline with 75k non-defaulters	76.8	76.0	0.569

Final Project: Discussion

- By re-balancing, we can earn the higher AUC but lose test accuracy a lot.
  - By decreasing samples of non-defaulters, we lose explanatory power for the test set.
- The baseline model without balancing shows the highest test accuracy, but the lowest AUC.
  - AUC represents the quality of model's predictions.
  - To beat the baseline model, we need similar test accuracy with higher AUC.
- Note that this is a baseline model without feature engineering and hyper-parameter tuning.
  - We have many options to enhance the baseline model, and also other candidate models.

Conclusion and Next Steps

- Our baseline model shows a great test accuracy but relatively poor AUC.
- Re-balancing is helpful improving the AUC at the cost of accuracy.
- We will start from considering other candidate models.
- Next steps (FP Phase 2) includes the processes
    - Additional Feature Engineering
    - Hyper-parameter Tuning
    - Feature Selection
    - Analysis of feature importance
    - Ensemble Methods