Technical Report

The dataset I used is this link: https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?select=Base.csv

It’s the base of the whole bank account fraud dataset suite.

This is a tabular dataset with 1 million instances and 31 features.

1. First, I did data understanding, and found there’s a column “device_fraud_count” just has one value for all instances, so I drop this attribute.
2. Then I checked if there are some attributes’ values are mostly missing. I found “prev_address_months_count”, “intended_balcon_amount”, so I drop these two attributes.
3. Then I impute the rest attributes with missing value. Some use -1 to represent missing values. Some use negative value as missing values. When impute numerical data, I use median. When impute categorical data, I use mode.
4. After imputation, I do train-test split based on attribute “month”, [0:5] as training, and [6:7] as test.
5. Because of the imbalance characteristic, I applied SMOTE oversampling techniques, and made two labels have equal quantity.
6. Then I did feature selection using domain and correlation.
7. After that, I did 1-in-100 systematic sampling.
8. After sampling, I used time-series validation.
9. To do modeling, I applied three techniques: Decision Tree, Random Forest, and Logistic Regression.
10. About measures, I use confusion matrix, Precision, Recall, F1-score, ROC_AUC, Matthew’s correlation coefficient to do comparison for effectiveness.
11. For Efficiency, I compared each model’s execution time.
12. For stability, I changed seed to 10, 500, 5000 to check the change of the metrics’ results.

Remaining work： Although I split the original data using “month” attribute. I compared different models based on month: [0:5], This technology divides the training data into train and test. So further work may be deploying model to original test data to check if they still work.