SUS 2025 BRUTTI MANZONI

Brutti Manzoni: Francesco De Martino, Roberto Magno Mazzotta, Beatrice Mazzocchi

Ingegneria Informatica, Automatica e Gestionale “Antonio Ruberti”

Data Science

Sapienza University of Rome

June 16, 2025

1. Problem Overview and Data Analysis

1.1 Task Description

The detection of fraudulent transactions (labeled as *"Is Laundering"*) presents a challenging classification problem in financial surveillance due to the rarity and subtlety of laundering behaviors. The model must distinguish between normal and suspicious patterns across diverse account and transaction characteristics.

1.2 Dataset Characteristics

A preliminary Exploratory Data Analysis (EDA) revealed no significant null values and identified only one self-referential transaction, which did not impede further analysis. The dataset contains 55,307 total transactions, with only 685 labeled as fraudulent (1.24%), presenting a severe class imbalance. This imbalance is a critical modeling concern: without mitigation, standard classifiers may favor the majority (legitimate) class, failing to detect rare but crucial fraudulent instances. Understanding this distribution guides us to adopt techniques like stratified sampling in later stages.

1.3 Key Findings from PCA

To identify the most influential variables on which to focus, we applied Principal Component Analysis (PCA). PCA revealed that only a few features account for the majority of data variance, highlighting their explanatory power.

Table 1: PCA Feature Importance (Explained Variance= 0.75)

Feature	**Importance**
Avg Stock Account To	0.500175
Avg Stock Account From	0.500000
Transaction count	0.499826
Amount Paid	0.500000

1.5 Data Preparation and Augmentation

Following the PCA, we proceeded with further data preparation and augmentation steps to address identified challenges, particularly the class imbalance.

1.5.1 Density Distribution Calculation

We calculated the density distributions for the key features identified by PCA ('Amount Paid', 'Avg Stock Account From', 'Avg Stock Account To') specifically for the fraudulent transactions. This allowed us to characterize the probability density functions of these features within the minority class, providing a more granular understanding of their typical ranges and patterns in fraudulent activities. We selected distributions that could accurately simulate the extremely left-skewed distribution observed in the raw data.

1.5.2 Synthetic Data Generation

To mitigate the severe class imbalance, we generated 2050 synthetic rows of data specifically for fraudulent transactions. This was achieved by sampling from the empirically derived density distributions of the fraudulent class for the critical features. This augmentation strategy ensures that the model has sufficient examples of fraudulent behavior to learn from, without simply oversampling the existing rare instances, thus reducing the risk of overfitting. After attempting to approximate critical features such as 'Avg Stock Account From/To' and 'Amount Paid', we arrived at the closest representation of the empirical distribution, which is the log-normal distribution. We obtained this result through a trial-and-error approach. This augmentation strategy ensures that the model has sufficient examples of fraudulent behavior to learn from, without simply oversampling the existing rare instances, thus reducing the risk of overfitting.

1.5.3 Feature Engineering (Transaction Count)

A new feature, 'Transaction Count', was engineered. This feature represents the number of transactions associated with a particular account, either as a sender or receiver. This aims to capture potential patterns where accounts involved in a high volume of transactions, especially within a short period, might be indicative of fraudulent activity. This feature was applied to both the training and test datasets.

1.5.4 Feature Selection and Scaling

Based on our analysis, including PCA, we removed columns deemed non-significant for the purpose of our analysis, focusing on features with high explanatory power. Subsequently, all numerical features in both the training and test datasets were scaled. This scaling process (e.g., using standardization or normalization) ensures that features with larger numerical ranges do not disproportionately influence the model's learning process.

2. Model Training and Evaluation

2.1 XGBoost Model Selection and Hyperparameter Tuning

For our classification task, we selected XGBoost (eXtreme Gradient Boosting) due to its robustness, efficiency, and proven performance in handling imbalanced datasets. To optimize the model's performance, we employed a Grid Search approach to tune its key hyperparameters. The primary metrics for evaluation were chosen to address the class imbalance, focusing on recall and precision for the minority class, alongside overall F1-score.

The hyperparameter grid for XGBoost included:

- **'eta': 0.05 (learning rate):** Controls the step size shrinkage to prevent overfitting.
- **'subsample': 0.8** The fraction of samples to be used for fitting the individual base learners. This helps to reduce variance.
- **'gamma': 4** Minimum loss reduction required to make a further partition on a leaf node. This parameter provides a pruning mechanism.

We performed 1000 iterations during the grid search and incorporated early stopping rounds (set to 20).

2.2 Validation on training Data

Once the optimal XGBoost model was identified through grid search, it was used to make predictions on the unseen test dataset. The predictions generated include both probability scores and binary class labels (1 for fraudulent, 0 for legitimate).