COMS 4995 Project Proposal: Fraudulent Transaction Detection

Haozhong Jia (hj2622), Nikhil Balwani (nb3096), Qingcheng Yu (qy2281), Zain Merchant (ztm2105)

1 Abstract

Financial fraud is a major issue faced by businesses and consumers alike, with nearly a $5.8 billion drain on our economy every year [1]. To fight against this and make transactions more safe, financial firms have increasingly been investing in automated tools and models to help predict the legitimacy of transactions and automatically block those that seem to be nefarious. This system is not perfect, however. There is a constant aim to reduce false positives that add additional friction and nuisance to the process, while increasing the accuracy of actual fraudulent charges that are blocked. Through our exploration of the IEEE-CIS Fraud Detection dataset [2], we hope to better understand the topic and apply some of the techniques learned throughout the semester towards this very important issue.

2 Data

The dataset is provided by Vesta Corporation from Kaggle. The data comes from Vesta's real-world e-commerce transaction. The dataset consists of two parts: identity and transaction. Two types of data are joined by the feature TransactionID that is a unique timedelta from a given reference datetime. The identity data contains features about identification information - network connection information and digital signature associated with transactions. Due to aspects of security and privacy protection, all the features are masked and represented by numerical id except Device types and Device systems. The transaction data contains features about money transfer and other consumption services. This information provides information about the purchase of the goods as well as information about the seller and seller. There are 871 columns in the dataset.

3 Methods

We plan to perform the following feature engineering steps:

1. Replace all missing values with an arbitrarily small value (-999)
2. Reduce memory usage by type-casting features
3. Normalize numerical columns

Subsequently, we will perform a train (60%), hold out (20%), and test split (20%) on the dataset. We will evaluate the following models on the validation set: Naive Decision Tree, Random Forest, XGBoost, and LightGBM. We choose tree based models because they are easy to understand and interpret. Finally, we will report the results of the best model on the test set. Due to an imbalance in the dataset, we will use Precision, Recall, F1 score, and AUC ROC to gauge our performance.

References

[1] CNBC Inc. *Consumers lost $5.8 billion to fraud last year — up 70% over 2020.* URL: https://www.cnbc.com/2022/02/22/consumers-lost-5point8-billion-to-fraud-last-year-up-70percent-over-2020.html.

[2] Kaggle Inc. *IEEE-CIS Fraud Detection.* URL: https://www.kaggle.com/competitions/ieee-fraud-detection/overview.