Customer Transactions Fraud Detection

Context

Vesta Corporation, the world's leading payment service company, is seeking the best solutions for fraud prevention. They build fraud prevention systems which save consumers millions of dollars per year. IEEE Computational Intelligence Society (IEEE-CIS) works across a variety of AI and machine learning areas, including deep neural networks, fuzzy systems, evolutionary computation, and swarm intelligence. Today they're partnering with Vesta, seeking the best solutions for the fraud detection industry. Researchers from the IEEE-CIS want to improve fraud detection accuracy and improve customer experiences. If successful, the model will improve the efficiency of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenues.

Problem Statement

What models can Vesta Corporation use to predict the probability that an online transaction is fraudulent, as denoted by the binary target isFraud?

Criteria for Success

The recommended model will accurately classify the fraudulent transaction over 80%, and reduce the false positive rate.

Constraints

1. The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement. It is unclear what some of the field names represent
2. The data is large, so the kernel may run slow during the interactive plots. It is possible and helpful to reduce the memory usage.

Stakeholders

The primary clients for this project are IEEE-CIS Researches, Vesta Corporation management team.

Data

1. The data comes from Vesta's real-world e-commerce transactions and contains a wide range of features from device type to product features, available in Kaggle competition (https://www.kaggle.com/c/ieee-fraud-detection/data)
2. The training datasets will be used in this project: train_transaction.csv (394 Columns) and train_identity.csv (41 Columns). The data is broken into two parts, identity and transaction, joined by TransactionID. Not all transactions have corresponding identity information.

3. Transaction datasets contains several categorical data: product code (ProductCD),, payment card information (Card1 - Card6), addresses (addr1 and addr2), card information matches (M1 - M9), P_ and R_ emaildomain
4. Identity datasets categorical data: DeviceType, DeviceInfo, id12 - id 38. Variables in the identity table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions. They're collected by Vesta’s fraud protection system and digital security partners.

Approaches

1. I will first filter the data, drop a few features with large missing values, or create new features
2. I will next explore the data, look for patterns, distributions, or relations of different features
3. Then, I will do some feature engineering make it easier for data training and computations
4. I will split the data into training set and test set. I will try a series of models on the training set, from basic logistic regressions to more advanced ones such as XGBoost, and compare models and choose the one based on f1, Precision, Recall rates
5. Once I pick the optimized model, I will predict the results on the test set

Deliverables

The final report will be a formal written report, with a slide deck for presentation purposes. Detailed steps and codes will be written in Jupyter Notebook.