Synthetic Financial Fraud Model

Problem statement formation

- How can we develop a fraud detection model that has a low false positive rate on synthetic data generated from mobile money transactions in Africa for use in a production environment immediately?

Context

- Since there is a lack of public financial data to develop fraud detection models, we need to see if it is possible to develop a highly precise model from a synthetically generated dataset

Criteria for success

- Extremely high precision rate

Scope of solution space

- Identifying what features contribute the most to identifying fraudulent money transfers

Constraints

- Synthetically generated dataset (from real transactions)
- Imbalanced dataset (low number of fraudulent labeled transactions)

Stakeholders

- Data scientist @ Fin-tech/Payments company

Data sources

- Kaggle - https://www.kaggle.com/datasets/ealaxi/paysim1

Assessing the risk of digital financial fraud is extremely difficult with the lack of publicly available data. What I am looking to accomplish is to see if I can successfully build a fraud detection model by utilizing a dataset that is synthetically generated from a mobile money service in an African country. The hope is that if this experiment is successful, it may provoke a response from data scientists from financial companies to offer more synthetically generated datasets to the public so that there can be an increased knowledge share of fraudulent trends.

Once retrieving the data sample, I will do some exploratory data analysis in order to determine if there are any data elements that need to be cleaned, dropped or even imputed before modeling. From there, I may look at generating my own features based on some of the correlations that I am seeing. Another thing that I will also have to consider is figuring out how to handle an imbalanced data set. An imbalanced data set in this case means that there is a lack of fraudulent data present – this may prove to be troublesome when trying to develop an extremely precise model for catching fraud. Potentially oversampling for fraud and undersampling legitimate transactions might be a way to overcome this problem.

Once the data is cleaned and features have been chosen, several models will be generated – XGBoost, Random Forest, Linear Regression are some that come to mind. Once the most optimal model has been discovered, a report will be generated as well as a presentation slide – each summarizing the findings from the modeling experiment.