Predicting Mobile Transaction Fraud Using XGBoost Ensemble Model

Written by: Jason Ortiz , Kshitiz Badola, Sobnom Mostari

Objective:

As technology advances it works in favor of both those trying to use it to exploit others and those using it to defend themselves. This has become the case for mobile money transactions, as new ways to steal information are available fraud has become a common issue to be dealt with. With the use of machine learning it is possible to create a model to help predict if a transaction is fraudulent or not, with the idea of being able to prevent a fraudulent transaction before it is too late. Our team plans to use a synthetic dataset found on Kaggle: "PS_20174392719_1491204439457_log.csv". This dataset is described as being a "Paysim synthetic dataset of mobile money transactions. Each step represents an hour of simulation. This dataset is scaled down 1/4 of the original dataset which is presented in the paper "PaySim: A financial mobile money simulator for fraud detection".". With this dataset we will be able to simulate production-like data, thus allowing us to test models on data that we would be likely to see if we entered the financial or banking industries. The team plans to take steps in analyzing the data through the EDA process, allowing us to see what preprocessing will need to be done before building any models. After ensuring the data is in a suitable format we intend to start with a few baseline models, likely K-Nearest-Neighbors, Random Forest Classifier, Logistic Regression, and possibly Gaussian or Bernoulli Naive Bayes classification models. After establishing a proper baseline for how our model should be performing, using metrics such as Accuracy, Precision, Recall, and AUC Curve, we can start training and cross validating our XGBoost Ensemble model. If the model doesn't outperform the baseline models we can potentially try adding XGBoost to a Voting Classification model and see if there's any improvement.