Executive Summary of Machine Learning-Based Detection of Credit Card Fraud

Credit card fraud presents a major source of risk to both businesses and customers. Accordingly, improvements in detecting fraudulent transactions present a business opportunity for credit card issuers. Using a dataset consisting of millions of transactions between a large variety of different consumers and merchants, dispersed across multiple different sectors and geographies, we trained and evaluated a machine learning model to flag fraudulent transactions.

We find that our machine learning method yields a marked increase in fraud detection: in terms of F1 score (a metric that summarizes the model's ability to predict both true positives and true negatives), the model achieves a score of 93%, as compared to a baseline value of 50%. Our final technique, the random forest, optimizes recall in order to minimize the false negative rate (i.e., failure to detect cases of fraud). The random forest has a recall rate of 89.6% and an F1 score of 93.0%. This indicates that our model correctly detects fraudulent transactions approximately 90% of the time. The top three features that drive the model's decision making, in decreasing order of importance, are transaction amount, distance between customer address and merchant address, and customer age. These features were found to have the greatest impact on fraud prediction in our model, with feature importance (usefulness in detecting fraudulent transactions) values of 0.73, 0.51 and 0.49 respectively.

The initial dataset consists of 41 million transactions from 999 distinct credit cards processed between January 1, 2019 and December 31, 2020. For each transaction, the dataset gives us various features, falling into three broad categories: (1) features relating to the merchant (including address and type of business), (2) features relating to the customer (including address of residence, basic demographic data, and employment), and (3) features relating to the transaction itself (amount, time, and credit card number). We first selected the transactions of 50 cards and reduced the sample size to 120,000 transactions. Because fraudulent transactions are so rare as a fraction of all transactions, we then up-sampled fraud instances to get a more balanced dataset to improve model training.

We constructed several new features: the card holder's age, distance between customer and merchant address, number of transactions from the same card in the last one hour period and the last twenty-four hour period, time of the day when the transaction happened, and indicators of whether the transaction amount is higher than certain thresholds. We also utilized the original variables of transaction amount, city population and customer gender, as features to train models. The techniques we tried include random guessing (as baseline), logistic regression, decision trees and random forest. We ultimately selected random forest which outperformed the others.