MLND Capstone Proposal

Grant Gasser, July 2019

Kaggle Competition: IEEE-CIS Fraud Detection

Domain Background

This project would be about detecting fraudulent transactions, which is a part of a larger field of research known as anomaly detection. This field is being transformed as computers get better and better at analyzing large datasets. See [Chandola et al 2009] for a general overview of anomaly detection techniques and applications.

Problem Statement

Credit card fraud costs consumers and businesses billions of dollars per year. As stated by Kaggle, creating a successful machine learning would "improve the efficacy of fraudulent transaction alerts for millions of people around the world, helping hundreds of thousands of businesses reduce their fraud loss and increase their revenue."

Datasets and Inputs

Dataset provided by Kaggle. The data include variables such as product names, card numbers, addresses, emails, device purchased from, and high dimensional transaction data. From sifting through the EDA kernels, there seems to be a significant amount of missing values.

Solution Statement

“In this competition you are predicting the probability that an online transaction is fraudulent, as denoted by the binary target isFraud.”

Benchmark Model

The leaderboard should provide a reference for the performance of my model. My goal is to place in the top 10% of the competition.

Evaluation Metrics

Area under the ROC curve, where the y-axis is the true positive rate, TPR = TP / (TP+FN) and the x-axis is the false positive rate, FPR = FP / (FP + TN).

Project Design

There will be a need for significant pre-processing of data since there are two train files and two test files. Both pairs need to be joined based on the transaction ID.

I would like to do some additional feature engineering which may include normalizing data, transforming variables (log or otherwise), creating interaction variables, or creating entirely new features.

Additionally, the transaction data is very high dimensional and will require PCA to reduce dimensions. I would try to follow the template outlined in the population segmentation project previously completed as part of the MLND.

There is also the issue with class imbalance. I will also follow the previous outline of how to balance (weight) the positive class more to account for imbalance.

Since the training data is very large, I plan on splitting the set into training and validation sets in order to try different models and hyperparameter tuning.