May 11, 2020

RE: ML Engineering Capstone Proposal

I am proposing to deploy a credit card fraud detection model using Tensorflow. The data is based on the well-known credit-card fraud dataset from a Kaggle competition run in 2018. The overview of the competition is on Kaggle at https://www.kaggle.com/mlg-ulb/creditcardfraud/.

Domain Background: Payment fraud losses reached nearly $28B worldwide in 2018, according to the Nilson Report, a financial services industry reporting service. While the US only accounted for ~20% global transaction volume in 2018, it represented 34% of total card fraud losses – making credit card fraud a real American problem. Most retail banking companies deploy machine learning models to identify fraud, as the volume of credit card transactions makes human ID of fraudulent transactions impossible.

Problem Statement: Identification of credit fraud presents a challenge to machine learning models however, as the data is often plagued with missing values. It is also necessarily highly balanced data – most financial transactions are legitimate, so any representative sample of transaction data will contain only a small fraction of truly fraudulent transactions.

Datasets and Inputs: The datasets for this competition are available on Kaggle via the link above. The test dataset contains ~280K financial transactions with roughly 30 features that have been masked to maintain the anonymity of the underlying data. Each observation also contains an amount, as well as a "No Fraud" or "Fraud" classification. The data is roughly 0.1% fraudulent, making the data imbalanced in the extreme.

Solution Statement: The goal of this project is to train a supervised model to accurately identify fraud based on the input data, with as a high a precision and recall as possible. The solution will likely require the synthetic creation of data in order to balance the dataset.

Benchmark Model: Several well documented approaches exist on Kaggle including the following: https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now That notebook achieves an AUC of ~0.95 with a logistic classification model, with under-sampling of the non-fraudulent data in order to manage the imbalanced nature of the data set.

Evaluation Metrics: Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) will be the primary metric used for evaluating the model quality. Precision and recall of a specific threshold as well as the classification confusion matrix will be used on test data in order to evaluate the quality of the solution.

Project Design: The Kaggle dataset covers 280K transactions, enough to separate the data into train and tests sets. Before modelling, techniques to synthesize fraudulent data will be used in order to create a more balanced training data set. The data is relatively clean, with few missing values, which should lend itself to modelling with a neural network approach. Extreme gradient boosting is another likely candidate for modelling.