Machine Learning Engineer Nanodegree

Capstone Proposal

Ricky J Sparks

April 28th, 2019

Domain Background

Fraud detection is a serious problem of the modern world. The news reports on countless attacks of credit card information being stolen annually. The transactions of fraudulent cards might seem minimal in occurrence but, on a larger scale these small occurrences can cost companies millions in losses. This is where machine learning comes into play learning from the patterns and adapting to new possible schemes. Machine learning can give credit card companies the ability to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Problem Statement

The problem presented is analyzing historical data of credit card transactions that were fraudulent and cards that weren't fraudulent. The goal of this model is to predict future transactions as fraud. The model will implement machine learning algorithms on the dataset that will ultimately lead us to an optimal algorithm that performs superior to the others. It is important that this model identifies most if not all of the fraud transactions with as much accuracy as possible. Thus, the nature of this problem is binary in nature as 0 represents valid transactions and 1 represents fraud transactions.

Datasets and Inputs

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

Dataset Source: https://www.kaggle.com/mlg-ulb/creditcardfraud

Solution Statement

The solution to this problem is using deployment of machine learning algorithms like Local Outlier Factor, Isolation Forest Algorithm. Thus, also using common metrics like precision, recall, and F-1 scores to give more clarity of these algorithms in measurements. More so, implementations of data visualization practices like correlation matrices and parameter histograms will also be used to provide more clarity of the data's distribution.

Benchmark Model

This dataset was obtained from Kaggle and based upon my observation the accuracy, f1-score, and recall shouldn't lesser than 75%. Otherwise, this would be a poor model falling short of other credit card fraud detection models in the data science space..

Evaluation Metrics

This model should be expected as a high recall model since the nature of the problem is to catch all fraudulent transactions including even small amount of transactions that weren't fraud. All in all, recall should be an important metric used to measure the data.

Project Design

The first step to approaching this project is retrieving a dataset of credit card fraud. The next step I would need to take is to install the necessary dependencies. Furthermore, then I would load the dataset and explore the data. From there derive the shape of the data and possibly apply other techniques to preprocess the data. Finally, once the dependencies are installed and the dataset is loaded correctly I can run machine learning algorithms. For example the Local Outlier Factor and Isolation Forest Algorithm. Thus, adjusting the parameters and the model architecture as whole could possibly lead to an optimal score.