Machine Learning Engineer Nanodegree

Capstone Proposal

Nitin Mahajan
June 17, 2018

Domain Background

Credit Card fraud losses are a huge problem. The Nilson Report, a publication covering global payment systems, reported recently that global card fraud losses equaled $22.8 billion in 2016, an increase of 4.4 percent over 2015. That amount does not include costs incurred by retailers, card issuers, and acquirers for their operations and chargeback management.

By 2021, card fraud worldwide is expected to reach a total of $32.96 billion. It is important that card companies can identify fraudulent card transactions so that customers are not charged for items that they did not purchase. Better tactics and technology to combat criminals attempting such transactions will define fraud fighting for merchants in coming years.

One of the main challenges while detecting frauds from financial transactions is to identify insignificant no. of fraudulent transactions amongst millions of non-fraudulent transactions. I'm working in the banking industry on solving similar problems where the dataset is highly unbalanced and that further motivation to work on this problem.

Problem Statement

The dataset for a credit card fraud is usually very unbalanced as the no. of fraudulent transactions are minimal as compared to total no. of transactions. This is a binary classification problem where the transactions are to be classified as fraudulent or not. As the classes are highly imbalanced, the accuracy would be measured using the Area Under the Precision-Recall Curve (AUPRC). Confusion matrix accuracy is not meaningful for unbalanced classification. The intuition behind using AUPRC is discussed in detail in the following sections.

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-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

This dataset has anonymized credit card transactions labelled as fraudulent or genuine and has been sourced from Kaggle (https://www.kaggle.com/mlg-ulb/creditcardfraud/data).

Solution Statement

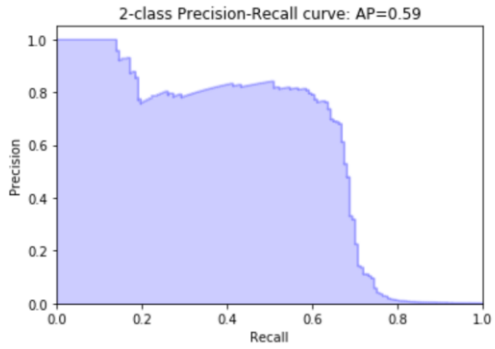
The main challenge with this classification problem is the highly imbalanced dataset. The fraudulent transactions are a very small fraction i.e. 0.172% of the overall data. In general machine learning models tend to maximize accuracy instead of recall. However, recall is our main concern as missing a fraud transaction would turn out to be very costly for a financial institution. Also, cross validation could be potentially biased when one or more folds of sample do not have any fraudulent transactions. In such a case our model will be very accurate and lead to deviation from the actual goal. Some techniques for assigning extra weight to fraudulent transactions & oversampling can be used for overcoming such challenges.

Benchmark Model

I created a benchmark model using logistic regression and used AUPRC(Area under Precision -Recall curve) as the measurement metrics. The average precision recall score was 0.59 for the model.

Evaluation Metrics

(approx. 1-2 paragraphs)



Given the imbalance ration, we would be using Area Under the Precision-Recall Curve (AUPRC) to measure accuracy. The closer to 1 the AUPRC is, the better the model is. A model with AUPRC score of 1 implies a perfect classifier.

Precision is the proportion of test cases predicted to be fraud that were indeed fraudulent (i.e. the true positive predictions), while recall or sensitivity is the proportion of fraud cases that were identified as fraud. This precision-recall curve(AUPRC) tells us the relationship

between correct fraud predictions and the proportion of fraud cases that were detected (e.g. if all or most fraud cases were identified, we also have many non-fraud cases predicted as fraud and vice versa).

F-beta score is can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0. So F-beta can be very useful while evaluating model performance. The F-beta score weights recall more than precision by a factor of beta. beta == 1.0 means recall and precision are equally important.

Project Design

I would start with exploring and understanding the data more.

Since the data is already encoded by PCA and variable description is not available we are somehow limited in the exploratory data analysis we can do.

Since all features are numerical, so we do not require any feature encoding or hashing. We would scale the data to have mean of 0 and variance of 1.

After scaling I would look at different options of over and/or under sampling the data to take care of the imbalance in the dataset.

Then I will split the dataset into training and testing data to apply different binary classification algorithms like Logistic regression and decision tree classifier. These would be evaluated against each other on the evaluation metrics to select the best model applicable.

After selecting the best model, further parameter tuning would be done to optimize results. The final optimized model would be used to classify fraudulent and non fraudulent transactions.