Machine Learning Engineering Nanodegree Capstone Proposal
Credit Card Anomaly and Fraud Detection
By Abhijeet Koranne
On 9th Nov, 2019

1. Domain Background: -

Financial fraud is ever growing menaces with far consequences in the financial industry. Payments are the most digitalized part of the financial industry, which makes them particularly vulnerable to digital fraudulent activities. Machine learning allows for creating algorithms that process large datasets with many variables and help find these hidden correlations between user behaviour and the likelihood of fraudulent actions. Anomaly detection is one of the common antifraud approaches in data science. It provides simple binary answers. It is based on classifying all objects in the available data into two groups: normal distribution and outliers. Outliers, in this case, are the objects (e.g. transactions) that deviate from normal ones and are considered potentially fraudulent. By analysing these parameters, anomaly detection algorithms can answer the following questions:

I. Do clients access services in an expected way?
II. Are user actions normal?
III. Are transactions typical?
IV. Are there any inconsistencies in the information provided by users?

As the quantitative nature of the financial domain and availability of large volumes of historical data make it significant area to apply ML. I am interested in implementing ML to financial business use case.
This project gives me the opportunity to tap the most common issue in banking sector.

2. Problem Statement: -

The Credit Card Fraud Detection Problem includes modelling past credit card transactions with the knowledge of the ones that turned out to be fraud. This model is then used to identify whether a new transaction is fraudulent or not.

3. Datasets and Inputs: -

The dataset is obtained from Machine Learning Group — ULB, Credit Card Fraud Detection (2018), Kaggle.[i]

The datasets contain transactions made by credit cards in September 2013 by European cardholders. These transactions are a subset of all online transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, where 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.

4. Solution Statement: -

There are various supervised machine learning algorithms that can be used independently or be combined to build more sophisticated anomaly detection algorithms.

5. Benchmark Model: -

Available dataset has 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, where the positive class (frauds) account for 0.172% of all transactions. Considering this imbalance, As Benchmark Model Area under Precision Recall curve score should have higher than 74 % score.

6. Evaluation Metrics: -

Usually performance of the model is evaluated based on confusion matrix with precision, Recall (sensitivity).

Precision = TP/ (TP+FP)
Precision gives the accuracy in cases classified as fraud (positive)

Sensitivity (or Recall) = TP/ (TP +FN)
Sensitivity (Recall) gives the accuracy on positive (fraud) cases classification.

But for given imbalanced class ratio in data set, I will measure accuracy using Area under Precision Recall Curve (AUPRC) score. The Area under Precision Recall Curve (AUPRC) is given by plotting Precision against Recall. A model with AUPRC score equal to one indicate perfect model.

7. Project Design: -

General Project steps are as follows: -

I. Data Visualization
   First the credit card dataset will be taken from the source and descriptive statistics of the dataset would be calculated to have a basic understanding of the distribution and structure of dataset.

II. Data Pre-processing and Evaluation
   Data cleaning and validation will be performed on the dataset which includes removal of redundancy, filling empty spaces in columns, converting necessary variable into factors or classes. then exploratory analysis would be carried out to understand how these variables contribute to the outcome.

III. Re sampling Data
   Considering the imbalanced dataset, the over Sampling Algorithm will be used to even out imbalance by constructing new points in minority class.

IV. Data Splitting
   Data will be divided into two part; one will be training dataset, and another will be test data set. Then K fold cross validation will be performed in which the original sample will be randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample will be retained as the validation data for testing the model, and the remaining k −1 sub sample will be used as training data.

V. Model Training
   Model will be picked from list of models like Logistic regression, Random Forest, Decision tree, SVM and Naive Bayes etc. Then Picked model will be trained and tested on dataset

VI. Model Evaluation

Then accuracy will be calculated based on with AUPRC score and a comparison will be made.

VII. Best scored model will be selected as solution.

References

i. Machine Learning Group — ULB, Credit Card Fraud Detection (2018), Kaggle

ii. **Nathalie Japkowicz, Learning from Imbalanced Data Sets: A Comparison of Various Strategies (2000), AAAI Technical Report WS-00–05**