Machine Learning Nanodegree

Capstone Proposal

Eswar Sesha Sai Kamineni

Date: 11th March,2019.

DOMAIN BACKGROUND

The PwC global economic crime survey of 2016 suggests that approximately 36% of organizations experienced economic crime. Therefore, there is definitely a need to solve the problem of credit card fraud detection. The task of fraud detection often boils down to outlier detection, in which a dataset is scanned through to find potential anomalies in the data. In the past, this was done by employees which checked all transactions manually. A lot of research has been done in order to find a solution to this problem. With the rise of machine learning, artificial intelligence, deep learning and other relevant fields of information technology, it becomes feasible to automate this process and to save some of the intensive amount of labor that is put into detecting credit card fraud.

PROBLEM STATEMENT

As already mentioned already that, with the help of machine learning techniques labor used in identifying the fraud transactions can be reduced to a great extent. Basically, the problem can be stated as a binary classification i.e. fraud transaction or genuine transaction. But the frequencies of the two classes is very unbalanced in this case, so, we don't have comparable number of observations for each classes. As part of this project, my aim is to create few machine learning models which can identify the fraud transactions from given data of transactions

DATASETS AND INPUTS

In order to reproduce this kind of problem, I found a useful dataset available on Kaggle . The datasets contain 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 and 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. Due to confidentiality issues, the authors 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-dependent cost-sensitive learning. Feature “Class” is the response variable and it takes value 1 in case of fraud and 0 otherwise

SOLUTION STATEMENT

To identify the fraud transaction, I will be implementing 3 different models and will compare their performances. I will start with a basic machine learning algorithm which is Logistic Regression. Logistic Regression is often used in problems with binary target variables. Our Class variable is indeed a binary variable. It is not the best approach, but at least it offers some insights in the data. Second model I will be implementing is using Random Forest. I’m planning to implement Random Forest model in 2 different ways. In first part I will be implementing the Random Forest algorithm overall data at once. But as the classes are highly unbalanced, in the second part I will be using under sampling technique (with different proportions) after dividing the whole dataset into 4 parts (let’s call them batches) and implementing Random Forest algorithm individually on each batch.

Lastly, I will be using an unsupervised technique named Autoencoders. The job of Autoencoder models is to predict the input, given that same input. I will be using reconstruction error involved in predicting the input again. I will use only genuine transactions for training the model and as a result the model should return high reconstruction errors for fraud transactions. In the end, I will compare all the models for their performance on highly unbalanced data.

BENCHMARK MODEL

As the problem is related to one of the most sensitive areas of implementation, individuals/organizationsusingmachinelearningtechniquesforsolvingthisproblemwhichincludesBanks, transaction agencies, etc. don’t usually share their works. Also, because the data is highly unbalanced, we can’t use Average Prediction as a benchmark model. I will be using the Logistic Regression Model as the benchmark model, an article related to which can be found at this article.

EVALUATION METRICS

Different aspects are important for the organizations using various techniques for identifying the frauds. Along with using machine learning techniques, most of the organizations use manual reviewsforthetransactionsreportedfraudbythesemodelsbecauseorganizationsdon’twanttoreport their genuine customers as fraud. Also, there is some cost involved with manual review of each transaction. So different aspects to consider in this problem are: • Total fraud transactions identified as fraud by the model (Recall). • Total genuine transactions (good customers) reported as frauds (False Positive Rate). • Total fraction of transactions marked as frauds and sent for manual review (True Positive+ False Positive Rate). Hence, I will be evaluating the different models based on these three statistics.

PROJECT DESIGN

As mentioned already, I will be implementing three different models for the given problem. More details for each of the implementations are as follows Logistic Regression: Using 70-30 test train split. Random Forest: I will be using multiple variances of Random Forest model which includes: – Using the whole data in single random forest with and without under sampling (90-10, 80-20, 70-30, 50-50). – Dividing the data into 4 different batches and training on 4 different random forests with and without under sampling (90-10, 80-20, 70-30, 50-50). • Autoencoder: I will remove the labels provided in dataset and use all the genuine transactions for training and calculate the reconstruction error involved in predicting the input. Fraud transactions will comparatively have higher reconstruction error then the genuine transactions. Based on the distribution of reconstruction error, I will choose a threshold above which I will mark transactions as fraud.

REFERENCES

https://www.kaggle.com/mlg-ulb/creditcardfraud/home
http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders
https://towardsdatascience.com/under-sampling-a-performance-booster-on-imbalanceddata-a79ff1559fab