Credit Card fraud Detection

Sravan Kumar Rachakonda
IIT BOMBAY
sravankumarrk14@gmail.com

***Abstract*—Credit card fraud detection is becoming really problematic now a days with increasing online purchases there are increasing Cyber attacks. Identifying them is very important. In this document we will go through how we did the analysis of a data set containing fraud and non fraud transactions and using different ML algorithms how are we able to predict the fraud cases.**

Keywords: Fraud detection, ML algorithms

I. Introduction

[1]Fraud can be committed in different ways and in many industries. The majority of detection methods combine a variety of fraud detection datasets to form a connected overview of both valid and non-valid payment data to make a decision. This decision must consider IP address, geolocation, device identification, “BIN” data, global latitude/longitude, historic transaction patterns, and the actual transaction information. In practice, this means that merchants and issuers deploy analytically based responses that use internal and external data to apply a set of business rules or analytical algorithms to detect fraud.

Credit Card Fraud Detection with Machine Learning is a process of data investigation by a Data Science team and the development of a model that will provide the best results in revealing and preventing fraudulent transactions. This is achieved through bringing together all meaningful features of card users’ transactions, such as Date, User Zone, Product Category, Amount, Provider, Client’s Behavioral Patterns, etc. The information is then run through a subtly trained model that finds patterns and rules so that it can classify whether a transaction is fraudulent or is legitimate. Now you know what fraud protection is, let’s look at the most common types of threats.

II. Data set

[2]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 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.

III. Analysis Pipeline

Using exploratory Data analysis I found that fraud cases have high amount withdrawn.

Firstly, there are really less occasions that is 0.6% fraud cases. So we need to upsample or down sample the data set so that there will be equal no. of data on both fraud and non fraud cases.

Upsampling: it increases the minority cases by repeating the cases so that they will be equal to majority, in our case it is non fraud

DownSampling: It decreases the majority cases so that they will be equal to minor cases, in this case data is extremely trimmed, so we may not get good results, so I negleted this option in our case

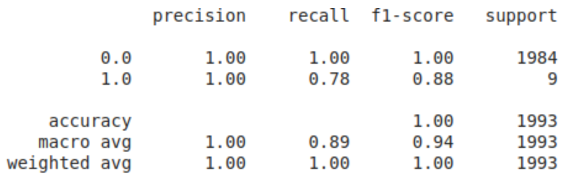
So I used a dummy predictor initially to see how better my models are in predicting fraud cases.

Next, I used Logistic regression it dosen’t give good results, but when I applied the Upsampled data, the results were a bit better in detecting the fraud cases.

Further I used Oversampler, which is more like a upsampler but it does better work than that. The random oversampler transform is defined to balance the minority class, then fit and apply to data set. The class distribution for transformed data set is reported showing that now the minority class has the same number of examples as the majority class[3]

Later i used random forest to the model, because The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced, and The random forest algorithm works well when you have both categorical and numerical features.

A screen shot from the Project of random forest predictions. It has a good recall and precision



```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00      1984
         1.0       1.00      0.78      0.88         9

    accuracy                           1.00      1993
   macro avg       1.00      0.89      0.94      1993
weighted avg       1.00      1.00      1.00      1993
```

CONCLUSION:

The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help.

References :

[1]credit card fraud detections solutions to business: https://spd.group/machine-learning/credit-card-fraud-detection/
[2]Credit card fraud detection: https://www.kaggle.com/mlg-ulb/creditcardfraud
[3]An article by Jason BrownLee: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/