Machine Learning Engineer Nanodegree

Capstone Proposal

Credit Card Fraud Detection Challenge

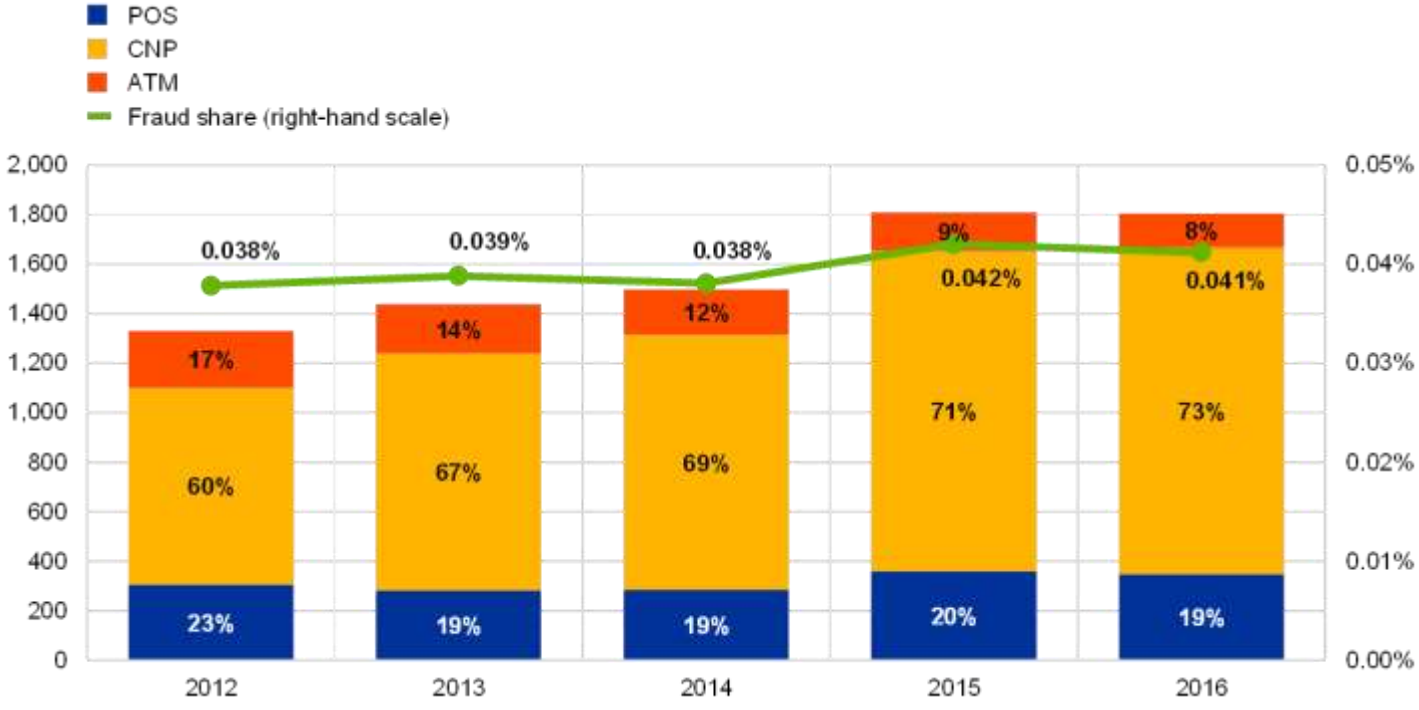
Rohit Venkatachalam
April 23, 2019

Proposal

Domain Background

Digital transactions are on the rise. And so are Credit Card and identity thefts. Below is a stats published by the European Central bank.

The image depicts Evolution of the total value of card fraud using cards issued within SEPA. (Single Euro Payment Account)



Source: All reporting CPSs.

Preventing fraudulent transactions is important, as it is a key component to promote digital transactions. Increase in digital transactions will lead to reduced cash transactions, and help in reduce crimes, arising out of cash transaction.

Problem Statement

Recognize fraudulent credit card transactions so that customers are not charged by credit card companies for items that they did not purchase.

Datasets and Inputs

The dataset provided in Kaggle, was collected by Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

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, the positive class (frauds) account for 0.172% of all transactions.

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.

Some of the past and future work in the related areas, can be viewed at
https://www.researchgate.net/project/Fraud-detection-with-machine-learning

Solution Statement

The solution is a classification model capable of predicting whether transaction is fraudulent of not.

I will use Python and Numpy to understand the data.

This is a highly imbalanced data, so will have to use a combination of over/under sampling.

I intend to use Deep Neural Network and it's adaptive learning capabilities to improve upon the existing fraud detection strategies.

Benchmark Model

As this is a binary classification problem. Logistic regression will be used to benchmark the results.Logistic regression gives 99% accuracy on this data. But this data is extremely unbalance. The number of fraud cases is miniscule. 492 frauds out of 284,807 transactions. To overcome this. Using random under sampling with Logistic regression gave an accuracy score of 93.9%

Evaluation Metrics.

The Kaggle data recommends measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC).

https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now

Project Design

Programming Language: Python

Library: Pandas, Numpy , Scikit-Learn, Keras

Workflow:

- Establish basic statistics and understanding of the dataset.
- Identify best approach to overcome the problem of imbalance in data. Techniques to be considered
    - Oversampling
    - Undersampling
    - Synthetic sampling(SMOTE)
- Train a base classification model
- Fine tune the model's hyperparameters.
- Perform training.

Architecture of the deep learning network. Classical Deep Neural Network

- I input layer with input dimension of 30
- 3 Hidden layers
- I output layer
- RELU activation
- Loss function is ADAM

References:

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

https://www.kaggle.com/gargmanish/how-to-handle-imbalance-data-study-in-detail

https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport201809.en.html

https://towardsdatascience.com/deep-learning-unbalanced-training-data-solve-it-like-this-6c528e9efea6