Capstone Proposal

Sharib Imam

Domain Background

Credit Card transactions has been a larger share of US payment system. In today's increasingly electronic society and with the rapid advances of electronic commerce on the Internet, the use of credit cards for purchases has become convenient and necessary.

Credit card transactions have become the de facto standard for Internet and Web based e-commerce. The US government estimates that credit cards accounted for approximately US $13 billion in Internet sales during 1998. This figure is expected to grow rapidly each year.

However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud. When banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. Hence, it is in both the banks and the cardholders' interest to reduce illegitimate use of credit cards by early fraud detection.

Credit Card companies are approaching data scientists in order find a better solution to this problem. To solve this problem, we need to build a model which flags the fraud transactions and gives an alert to the companies and the cardholders. This model can be designed using both supervised learning and unsupervised learning methods. For supervised learning methods, we need to have labelled data to train our algorithm. Whereas for unsupervised learning methods, we can tag the outlier transactions as fraud.

A research collaboration of Worldline and the Machine Learning Group of ULB on big data mining and fraud detection have worked on the datasets provided by Kaggle in order to find a better model to tag the fraud credit card transactions.

Academic Papers on Credit Card Fraud Detection using Supervised and Unsupervised Models

1) https://ieeexplore.ieee.org/document/5159014/

2) https://pdfs.semanticscholar.org/1752/a117dec81740c1d5516be15a3395a6d74a3c.pdf

3) http://ijcttjournal.org/Volume4/issue-7/IJCTT-V4I7P143.pdf

Problem statement

The Fraud Detection Problem includes modelling past 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. Our aim here is to detect as much of the fraudulent transactions as possible, while minimizing the incorrect fraud classifications.

Given the unbalanced nature of this dataset, we most likely need:

- to find a way to rescale/resample the dataset to be more balanced and in a way that allows us to approach the problem as a normal balanced classification problem, I will need to look into resampling methods/techniques that would best suit this problem, most probably it would be under-sampling, which deletes instances from the over- represented class in order to find an almost 50/50 representation for both classes.

- After that, I will test some supervised learning algorithms as well as unsupervised algorithm to find the best possible algorithm, while not consuming much computation power.

Datasets and Inputs

For this project, I am going to use the datasets which contains transactions made by credit cards in September 2013 by European cardholders which are available on Kaggle. The dataset contains 492 frauds out of 284,807 transactions that occurred in 2 days. Due to confidentiality issues, the dataset contains only numerical input variables which are the result of a PCA transformation. It consists 28 Principal Components, along with Time and Amount of the Transaction.

Since, we have only 492 frauds, the dataset is considered as class imbalanced. In order to overcome this, we have to perform under sampling or up sampling on the dataset to make it balanced.

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 dependant cost- sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. The remaining 28 Principal components gives us the transformed personal information about the cardholder.