Project Proposal: INFO 7390 Advances in Data Sciences
Credit Card Fraud Detection

Overview

Credit card fraud detection relies on the analysis of recorded transactions. Transaction data are mainly composed of several attributes (e.g. credit card identifier, transaction date, recipient, amount of the transaction etc.). Automatic systems are essential since it is not always possible or easy for a human analyst to detect fraudulent patterns in transaction datasets, often characterized by many samples, many dimensions and online updates. Also, the cardholder is not reliable in reporting the theft, loss or fraudulent use of a card which may become a problem later for both parties. We propose to use different machine learning classification algorithms to decipher a fraudulent transaction from a genuine one.

Dataset

We are using the Credit Card Fraud Detection dataset from https://data.world/vlad/credit-card-fraud-detection/discuss/credit-card-fraud-detection/mm4wiyjv
The data set contains 32 columns.
- Column_a: This column is a serial number for all the transactions in the dataset
- Time: Time depicts the number of seconds from the first transaction in this dataset
- V1-V28: Upon doing some research the owners of the data set did not want to make its content public. Hence, they performed PCA (Principal Component Analysis) on the features and then published the results.
- Amount: Amount of the transaction
- Class: Variable indicating if the transaction is fraudulent or genuine

Approach to the problem

We performed some analysis on the data set and noticed that only 492 out 284,807 are true. This is a clear case of uneven distribution of data which will not help us solve our problem. Therefore, we will be performing under sampling and over sampling over our data to ensure we can perform proper classification on both our resulting datasets.
- Perform feature engineering on datasets to select best features for our classification problem
- Apply various classification models (Logistic Regression, Random Forest, Support Vector Machine, Naïve base Classifier)
- Decide the best hyper parameters for the above algorithms
- Perform cross validation and evaluate confusion matrix and dice matrix to determine accuracy of the model
-Perform different autoML techniques to evaluate any other unexplored algorithms
-Present final pipeline

Deployment

We will deploy our model in the form of a web application using flask, so it can be accessed without any dependencies on other machines.
Additionally, we'll give the users parameters for a year range. The output would be show our predictions and the true results.