Springboard Data Science Career Track

Capstone Two - Project Proposal

Detecting credit card fraudulent transaction

Presented by: Hema Saxena
June 2020

● Problem statement formation

Given past credit card transactions of European cardholders, with the knowledge of the ones that turned out to be fraud, which new transaction can be marked as fraudulent?

● Context

Fraud is a major problem for the whole credit card industry that grows bigger with the increasing popularity of electronic money transfers. To effectively prevent the criminal actions that lead to the leakage of bank account information leak, skimming, counterfeit credit cards, the theft of billions of dollars annually, and the loss of reputation and customer loyalty, credit card issuers should consider the implementation of advanced Credit Card Fraud Prevention and Detection methods. Machine Learning-based methods can continuously improve the accuracy of fraud prevention based on information about each cardholder's behavior.

● Criteria for success

1. Supervised and unsupervised models will be built to estimate the probability of a new transaction to be fraudulent or not.

2. The performance of these models will be evaluated with respect to appropriate metrics, in alignment with the business problem.

3. Models will be compared with respect to these performance metrics, and recommendations will be offered to the client using the results of this comparison.

Scope of solution space

These transactions were made by European cardholders and so this project will only analyze behavior of European cardholders.

Not every fraudulent transaction is caught and reported, so any mis-classified data cannot be worked upon/handled in this project.

● Constraints

Analysis will be done using the dataset that presents transactions that occurred in only two days, with 492 frauds out of 284,807 transactions which is only 0.172% of all transactions. If more data points could be added, the models can make better predictions.

● Stakeholders

Credit card issuing organizations like banks and financial institutions.

● Data sources

CSV File available from Kaggle

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-sensitive learning.
Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

● Anticipated Solution Approach(es)

To model the business problem, I anticipate using

1. Supervised learning using the following algorithms:
    a. Logistic Regression
    b. Random Forests

2. Semi-supervised way for anomaly detection using the following algorithms:
    - Isolation Forest Algorithm
    - Local Outlier Factor (LOF) Algorithm
    - One-Class Support Vector Machine

To evaluate the performance of both the above data science approaches I currently envision, I will be evaluating the following metrics **per class.** (For a Semi-supervised approach, I will drop the labels first for predicting the outputs and then bring back the labels to train the models to predict the fraudulent transactions. Finally, I will evaluate the performance using the below methods)

Accuracy: To check the fraction of predictions that the model gets right for a given class.

Recall: Out of the fraudulent transactions, what percentage of these are correctly identified by the model for a given class?

Precision: Out of all the transactions predicted to be fraudulent, what percentage were fraudulent, for a given class?

F1 Score: Determine the performance of the model by combining Recall and Precision into one metric.

● Deliverables:

A GitHub repo containing the work done for each step of the project, including:
■ A slide deck
■ A project report
■ Jupyter Notebooks