# Machine Learning Engineer Nanodegree

## Capstone Proposal - Credit Card Fraud Detection

By -

Amber Ved

www.github.com/amberved

May 1st, 2017

# Proposal

## Domain Background

Big and high profile Credit cards and data breaches have been dominating the headlines in the past couple of years across the world. These problem of Credit cards breaches in U.S. alone is responsible for 47 percent of the world’s card fraud as of 2014. 15.4 million US consumers were affected by these kind of fraud in 2016, which is nearly 2 million more than in 2015. In dollar amount associated with such activities is 16 Billion in 2016 alone with ever Significantly increasing Card-Not-Present Fraud by 61 Percent. Hence it is a big clear and present problem that needs Smarter soultions and machine learning can help.

# Problem Statement

According to data from the Federal Reserve, Credit card Fraud only impacts a fraction of all purchases made with Credit Cards but it represents one of the biggest concerns among consumers and also results into billions of dollars of losses to fininial companies be it bank, credit card companies, retailers & governments. This is a complex problem with large number of marid financial aspects. Consequently we need automatic systems able to support fraud detection and fightback. These 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 a large number of samples, many dimensions and online update.

# Datasets and Inputs

I am planning to use Kaggel dataset on “Credit Card Fraud Detection” found at https://www.kaggle.com/dalpozz/creditcardfraud

This dataset presents Credit Card transactions, where it has 492 frauds out of 284,807 transactions. It contains around 30 features for each transection which are PCA transformation. All Features V1, V2, … V28 are the principal components obtained with PCA. Only 2 features which have not been transformed with PCA are ‘Time’ and ‘Amount’.

The intereseting aspect about using this data set is that ratio of frauds vs total transactions is very low. Secountly, all the data is already

convered into PCA so we can not judge the importance of any feature over other and have will force us to work with all of them equally and focus on the r2 or similar mathamatical relationships in place of human intution. Last point about this data set is that this is really good dataset as data preprocessing and cleansing is already done and with few changes can be fed into many supervised learning algorithms.

## Solution Statement

In this section, clearly describe a solution to the problem. The solution should be applicable to the project domain and appropriate for the dataset(s) or input(s) given. Additionally, describe the solution thoroughly such that it is clear that the solution is quantifiable (the solution can be expressed in mathematical or logical terms) , measurable (the solution can be measured by some metric and clearly observed), and replicable (the solution can be reproduced and occurs more than once).

## Benchmark Model

This problem of identifying frudulant transection from valid credit card, can we taken as a classical ML Cluster and/or Classfication problem.

Cluster in general is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. Since this problem have 4 possible conditions for any transection

Identified as Fraud and is flagged as Fraud
Identified as Fraud and is *NOT* flagged as Fraud
Identified as *NOT* Fraud and is flagged as Fraud
Identified as *NOT* Fraud and is *NOT* as Fraud
Hence can try to use confusion matrix to review outcome of the ML models.

I will prefer to use various clustering algorithm like K-means clustering algorithm, Gaussian Mixture Model clustering algorithm & along with other classficition models like decisionTree or Randomforest.

## Evaluation Metrics

As outlined I think confusion matrix could be a good way to evalute various models and algorithum effectiveness. Specially one can focus on having mild bias towards over reporting on -ve side (detecting transections as fraud in place of more relax where fraud transection are under reported). This could be better overall strategy as unreported fraud transection can result into financial loss. We dont have to different data sets for developing and benchmarking hence we will use random cross validation to compare various models. I will also like to benchmark our outcome against various other implementation avaliable on Kaggle.

Last but not the least, I will share the details of my model on Kaggel for other to review and share insight.

# Project Design

I intentd to follow the strategy to approach this problem:

A) Explore the dataset - Understand the data by finding basic relationships in data between various features and target variable.
B) Do necessary data preprocessing like feature reduction etc and use cross-validation to train several different ML algorithms (K-means clustering algorithm, Gaussian Mixture Model & others). Get a baseline score and confusion matrix.
C) Tune each model and validate noticable score improvements for each of them.
D) Train an ensemble using the stacking technique with previous models as base predictors. To get the final score.

---

## cited and references


https://www.javelinstrategy.com/coverage-area/2017-identity-fraud

https://www.businesswire.com/news/home/20170201005166/en/Identity-Fraud-Hits-Record-High-15.4-Million

https://www.kaggle.com/dalpozz/creditcardfraud

http://blog.kaggle.com/2016/07/21/approaching-almost-any-machine-learning-problem-abhishek-thakur/

https://en.wikipedia.org/wiki/K-means_clustering

https://en.wikipedia.org/wiki/Mixture_model

http://www.cs.cmu.edu/~guestrin/Class/10701-S07/Slides/clustering.pdf