# Credit Card Fraud Detection

# Machine Learning Capstone Project

Proposed by: Hafizullah Mahmudi

Date: Saturday, September 7, 2020

## Background

Credit cards and data breaches over several years have made fraud a big concern for many people.

Four out of five of the largest data breaches ever recorded occurred in 2016 which affected over 1 billion people.

A data breach in 2017 affected over 3 billion people. 5.8 occurred in banking, credit and financial sector.

According to the Federal Trade Commission's online database of consumer complaints has compiled 13 million complaints from 2012 to 2016, with 3 million in 2016 alone. Of those, 42 percent were fraud related, and 13 percent were identity theft complaints.

About 1.3 million complaints were fraud-related. Consumers reported paying over $744 million in those fraud complaints; the median amount paid was $450. Fifty-one percent of the consumers who reported a fraud-related complaint also reported an amount paid.

The US credit card balance has hit over $1.0645 trillion by April 2019.

## Problem Statement

The purpose of this project is built and test machine learning models to find if a transaction is fraudulent or not in the available data set that has already been labelled fraudulent for few transactions. We will employ machine learning techniques to detect fraudulent transactions as much as possible while trying to cut down the classification errors.

## Datasets and Inputs

The dataset that will be used with this project contains real -world transactions made through credit cards by European cardholders in September 2013 which is available in Kaggle.

This dataset presents transactions that occurred within 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.

It 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 there 492 frauds out of

284807 the data is highly imbalanced. The PCA features are labeled as V1, V2, …. V28 with Time and Amount in original form.

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

## Understanding and Cleaning Data

We will load the data into jupyternotebook and understand the available features and types. Then clean the data if there is any outlier, null values and invalid data.

## Exploratory Data Analysis (EDA)

I will use different visualization technique to perform univariate and bivariate analysis and transform time and amount transformation if required. Look at the skewness of the data and mitigate.

## Train/Test Split

After EDA, we will split the data into train and test split. Here, for validation, I will use the k-fold cross-validation method. I will choose an appropriate k value so that the minority class is correctly represented in the test folds.

## Proposed Solution

Since the data is imbalanced, we will use ADAptive SYNthetic (ADASYN) or SMOTE based on time availability and feasibility to balance the data and then build models and extract the results.

## Model Building/Hyperparameter Tuning

Looking at the other scores achieved from other Kaggle contributors, I will try to use different model to achieve over 80% recall, f1-score and the accuracy.  Will employ hyperparameter tuning until the desired level of performance on the given dataset is achieved.
I will try random forest and xgboost if there is enough time and resources available so that I can run these machine learning techniques with proper hyperparameter tuning.

## Evaluation Metrics

Since the data is imbalanced, we recommend measuring the accuracy using the Area Under the Precision-Recall Curve (AUPRC) and other metrics like accuracy, recall, and f1-score. A higher recall (the

ratio of correctly predicted positive observations) should be a good metric to catch suspicious transitions that were wrongly classified as normal transaction.

## References:


1. https://www.kaggle.com/mlg-ulb/creditcardfraud
2. https://www.creditcards.com/credit-card-news/credit-card-security-id-theft-fraud-statistics-1276/
3. https://www.creditcards.com/credit-card-news/credit-card-debt-statistics-1276/
4. https://ieeexplore.ieee.org/document/4633969