## Credit Card Fraud Detection

## Problem Statement

In Banking domain, it is important to identify fraudulent transactions as it not only penalizes the bank but also results in dissatisfaction from highly profitable customers, making them churn. Credit card Fraud happens quite often with the digital payment channels. To prevent such fraudulent transaction’s Machine Learning can be used to perform reviews of transactions which are costly and time consuming on human reviews and denying transactions which are not legitimate.

## Benchmark Model

I plan to compare my models with Basic Logistic Regression Classifier as my bench mark model, since it is considered as a go to method for binary classification problems. I want to fit Logistic Classifier to the labelled dataset and calculate the Area under the Precision- Recall Curve and Recall scores. I would like to choose these performance metrics as a threshold to evaluate my predictive models.

## Evaluation Metrics

Since our target is to catch the fraudulent transactions, and hence we would rather wrongly catch suspicious transactions than to mistake fraudulent transactions as normal

ones, then Recall (the ratio of correctly predicted positive observations to all the observations in actual class.) along should be our main metric. Yet, we will still aim to have a somewhat accurate model as much as possible.

## Problem Solution

The dataset is PCA transformed except for time and amount columns. Because of newly obtained Principal Components the interpretability of features in the dataset is lost.

There are no missing values in the dataset.

There is need to check for skewness of the features, as it can mess up with the model’s predictive power, which will be corrected using Log Transformation

**Class Imbalance (Minority Class Problem):**

The dataset presents transactions with 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

Both SMOTE (Synthetic Minority Over-Sampling Technique) and ADASYN (Adaptive Synthetic) can be used for the purpose, but for unbiased result ADASYN will be chosen.

**Algorithms and Techniques**:

After all the above-mentioned data pre-processing steps, choose the predictive models to perform the predictions and check for their ROC-AUC and their Precision.

**Logistic Regression**: its basic Algorithm and it works only linearly separable data and its one of the pros. If there is overlap in the data, then logistic regression fails and easily understandable algorithm and highly interpretable.

**KNN:** K-nearest neighbor is a simple, supervised machine learning algorithm used for both classification and regression tasks. It performs these tasks by identifying the neighbors that are nearest to a data point. For classification tasks, it takes the majority vote and for regression tasks, it takes the average value from the neighbors.

**Decision Trees:** The main cons are we really don’t know when to stop the tree training and when the tree id over fit. The biggest pro’s they give interpretation in flowcharts, for interpretation views it’s almost equal to Logistic regression.

**XGBoost**: stands for extreme Gradient Boosting.it takes the view of loss function. We need to build first tree to reduce the loss and second tree will reduce more loss and so on. And each new tree tries to reduce the loss and we have loss to zero which is called Gradient Boosting. The reason it is called extreme as it takes L2 Regularization.

**Hyperparameter Tuning**:

By using broad range models, we will reach very good accuracy (AUC). To get AUC hyper parameter tuning plays vital role

**Train-Test and validation:** If the quantity of data is very high and we have enough amount of data for all three Train, test and validation. If the data is very less like our Fraud test data we have only 500 records so it’s very difficult for this method

**K-Fold Cross Validation:** Evaluating the performance when the data set is randomly split into ‘k’ groups. Out of these groups, one will be used as the test set and the rest of the groups will be used as train sets. To evaluate the performance, the model will be trained on k-1 groups and then scored using the test set. This process will be iterated until each unique group has been used as the test set.

**Stratified K-Fold Cross Validation:** we rearrange the data to ensure that each fold is a good representative of all the strata of the data. Stratification ensures that each fold is representative of all the strata of the data. The class ratio will be reflection

Grid-Search: instead of training of each and every tree we are training 10 set of trees so we get broad area which tree is performing good and it uniformly sample. In real time scenario it’s good to have grid search as we uniformly sample

**RandomiszedSearchCV:** it will randomly pick random samples of pick for. If there are maximum nodes it’s good to try randomized search cv

scikit-learn in the form of RandomiszedSearchCV and GridSearchCV.

Compare the models, against their ROC-AUC and their Precision and finally conclude on which predictive model performs best.