**Credit Card Fraud Detection**

**Problem Statement:**

The aim of the project is to predict the fraudulent credit card transactions using machine learning models. The data set contains thousands of individual transactions that took place over a course of 2 days and their respective labels. These are the transactions made by European card holders over a period of two days in September 2013. Out of 284807 transactions, only 492 were fraudulent. Thus, the data seems to be highly imbalanced. Also, the data set has been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from time and amount, all other features are obtained by PCA.

**Planned Approach:**

1. Data Understanding: Load the data and get an understanding of the features.
2. Exploratory Data Analysis (EDA): Univariate, bivariate analysis to be done. Check for any skewness in the data.
3. Train/Test Split: Here we check on the model performance on unseen data. For validation k-fold cross validation to be used.
4. Model building/ Hyper parameter tuning: Here different models can be used to check on the performance by fine tuning the respective models hyperparameters. **Stratified K-Fold Cross Validation** is an extension of K-Fold cross-validation,in which we rearrange the data to ensure that each fold is a good representative of all the strata of the data.

**Class Imbalance:**

Before splitting the data into train/test set, the class imbalance must be handled. Here SMOTE (Synthetic Minority Over-Sampling Technique) algorithm can be used. This is a technique used to oversample an imbalanced data set adding synthetic points. **ADASYN** is similar to SMOTE, with a minor change i.e. the number of synthetic samples that it will add will have a **density distribution**. The aim here is to create synthetic data for minority examples that are harder to learn, rather than the easier ones.

**Models:**

The below models can be checked to predict the credit card fraudulent transactions. Finally, can conclude on the model that provides better accuracy.

1. Random Forest Model
2. XGBoost (eXtreme Gradient Boosting)
3. KNN (K Nearest Neighbors)

Since the data seems to be structured, the above models tend to perform better. Out of the above the model that gets to best accuracy can be showcased as the model of our choice.

**Model Evaluation:**

Accuracy is not always the correct metric for solving classification problems of imbalanced data. Because the ROC curve is measured at all thresholds, the best threshold would be one at which the **TPR is high and FPR is low,** **i.e., misclassifications are low.** Depending on the use case, we will have to account for what we require: high precision or high recall.