Credit Card Fraud Detection

Problem Statement

- It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.
- Retaining high profitable customers is the number one business goal.
- Base on the report by Nilson, banking frauds would account to $30 billion worldwide by 2020.

Objective

- To predict fraudulent credit card transactions with the help of machine learning models

Project Pipeline

- project pipeline can be briefly summarized approach for solving above problem.

Understanding Data

- Data set includes credit card transactions made by European cardholders.
- Data is highly unbalanced, **with the positive class (frauds) accounting for 0.172% (492 fraudulent) of the total transactions (2,84,807).**
- Data set has been modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from 'time' and 'amount', all the other features **(V1, V2, V3, up to V28)** are the principal components obtained using PCA.
- The feature 'time' contains the seconds elapsed between the first transaction in the data set & the subsequent transactions. The feature 'amount' is the transaction amount.
- The **feature 'class' represents class labelling**, and it takes the value 1 in cases of fraud and 0 in others.
- Based on the dataset it is a **"Minority Class Problem**" where feature classes are unequally divided.
- Methods to mitigate "M**inority class problem"** problem
  - Undersampling
  - Oversampling
  - Synthetic Minority Over-Sampling Technique (SMOTE)
  - ADAptive SYNthetic (ADASYN)

Exploratory data analytics (EDA)

- Current data set uses Gaussian variable, no need to perform Z-scaling
- However, need to check **skewness** in the data and try to mitigate it, as it might cause problems during the model-building phase.
- skewness may affect model assumptions or may impair the interpretation of feature importance.

Train/Test Split

- Train/test split can help to check the performance of your models with unseen data.
- We can use the **Stratified K-Fold Cross Validation** method as data is imbalanced, we need to choose an appropriate k value so that the minority class is correctly represented in the test folds.
- **Stratification** ensures that each fold is representative of all the strata of the data.

Model-Building/Hyperparameter Tuning

- Credit card fraud detection is a classification problem
- Selection of different models based on the **type of data & Model building** like data is linearly separable and need to be interpretable then logistic regression fits best.
- Different Models
    - KNN
    - Decision Tree
    - Random Forest
    - Gradient Boosting
    - XBoost
    - Deep Neutral networks
- Hyperparameter tuning settings controls models. Ideal settings of a model used for a particular data set will differ from those of models used for other data sets.
- We need to fine-tune their hyperparameters until you get the desired level of performance
- **Grid Search** can be thought of as an exhaustive search of hyperparameters for selecting the ideal hyperparameters for a model.
- We can use iterative method to select hyperparameters.
    - First select a nearby range on which the model might perform well.
    - Next, we will look at more samples within that range to find the best value within that grid

Model Evaluation

- Model evaluation helps to find the best model that tells how well the chosen model will work in the future.
- Model evaluation help us to avoid overfitting problem.
- Accuracy is not always the correct metric for solving classification problems.
- There are other metrics such as precision, recall, confusion matrix, F1 score, and the AUC-ROC score
- The **ROC curve** is used to understand the strength of the model by evaluating the performance of the model at all the classification thresholds.
- For **banks with smaller average transaction value**, we would want high precision because we only want to label relevant transactions as fraudulent.
- For **banks having a larger transaction value**, if the recall is low, i.e., it is unable to detect transactions that are labelled as non-fraudulent

The above is approach for solving the above-mentioned problem. The final implementation though might differ.