Capstone Project Proposal: Credit Card Fraud Detection

For many banks, retaining high profitable customers is the number one business goal. Banking fraud, however, poses a significant threat to this goal for different banks. In terms of substantial financial losses, trust and credibility, this is a concerning issue to both banks and customers alike. With the rise in digital payment channels, the number of fraudulent transactions is also increasing with new and different ways. In the banking industry, credit card fraud detection using machine learning is not just a trend but also a necessity for them to put proactive monitoring and fraud prevention mechanisms in place. Machine learning will help these institutions to reduce time-consuming manual reviews, costly chargebacks and fees, and denials of legitimate transactions.

The dataset used in this project is from Kaggle platform. The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. Out of **284,807** transactions**, 492** were fraudulent**.** This data set is highly unbalanced, with the positive class (frauds) accounting for just **0.172%** of the total transactions**.** The data set modified with Principal Component Analysis (PCA) to maintain confidentiality. Apart from **'time'** and **'amount'**, all the other features **(V1, V2 and 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 and 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.

The aim of this project is **to predict fraudulent credit card transactions using machine learning models.** Therefore, we will begin with first exploring the raw data to check for missing values and incorrect entries. In this dataset, we have only two columns with raw data; we can explore the two columns with scatter plot to observe the distribution with the target variable. After observing the distribution of principal components through histogram, the data will be observed if normally distributed which is expected after PCA transformation, but if some features are skewed, will be handled by applying power transform package from sklearn module to make data more Gaussian-like. As per the data description, we know that the class distribution is highly imbalanced, we will handle by both SMOTE and ADASYN techniques and build model on using both. The data processing step will split the data into train-test split in the ratio 80:20. Next, we will build the model, first with KNN, followed by Random Forest and XG-Boost machine learning algorithms. We will perform coarse hyper-parameter tuning for different hyper-parameters using GridSearchCV package followed by fine hyper-parameter tuning on the resulted optimum parameters to get final optimal parameters. For all models, we will perform the tuning in the same approach. In addition, for every model we will evaluate the performance using the Area Under the ROC Curve (AUC score) because confusion matrix accuracy is not meaningful for unbalanced classification.

The project is targeted to achieve the accuracy of more than **95%** to **predict the frauds in real-time** for saving millions of dollars from the billion dollar credit card company.