Credit Card Fraud Detection – Viva

- **Problem Description:**

Credit card fraud is a major threat for worldwide financial institutions. The aim of this Capstone Project on Credit Card Fraud Detection is to build a machine learning model capable of detecting fraudulent transactions.

The project pipeline can be briefly summarized in the following five steps:

- **Data Understanding:**

The data set includes credit card transactions made by European cardholders over a period of two days in September 2013. This data set is highly unbalanced, with the positive class (frauds)of about 0.172% of the total transactions. Hence appropriate Data Imbalance Handling Techniques such as SMOTE or ADASYN have to be used

- **Exploratory data analytics (EDA):**

Univariate and Bivariate Analysis could be performed on the dataset in this step. Apart from ‘time’ and ‘amount’, all the other features are the PCA transformed to maintain confidentiality. Hence the distribution plots of the variables were Gaussian. So, Z-scaling is not needed. But skewness of the data has to be checked for and mitigated.

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.

- **Train/Test Split:**

Train/Test split is performed on the Data to check the performance of the model on unseen data. For validation Stratified k-fold cross validation method could be used since the data is highly imbalanced. Since there is data imbalance k-value have to be appropriately chosen so as to represent minority class in the test folds

- **Model-Building/Hyperparameter Tuning:**

We need to create different models like Logistic Regression, Decision Tree, Random Forest, XGBoost etc and apply the boosting techniques to improve the model performance. The hyper parameter of each model should be tuned to get the best performance out of the model. Random and Grid Search methods could be used for hyper parameter tuning.

- **Model Evaluation:**

Since the data set is highly imbalanced precision, recall, confusion matrix, F1 score etc cannot be used to measure the model performance as these metrices are dependent on the threshold. The best threshold would be one at which the TPR is high and FPR is low, i.e., misclassifications are low. We use an AUC/ ROC curve or AUCROC score to determine the performance of the model. To save banks from high-value fraudulent transactions, we have to focus on a high recall in order to detect actual fraudulent transactions.