Problem Statement-

The problem we have to solve is to predict fraudulent credit card transactions with the help of machine learning models.

About the Data Set-

- **The data set provided i**ncludes credit card transactions made by European cardholders over a period of two days in September 2013.
- The 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 'class' represents class labelling, and it takes the value 1 in cases of fraud and 0 in others.
- We have a highly imbalanced Class which needs to be handled before creating models for predictions.

Approach to solving the problem-

- **Load data for understanding**- Load the data and understand the features present in it. This would help in choosing the features that will be required for the final model.
- Perform **Exploratory data analysis (EDA)-** As Gaussian variables are used in the dataset, instead of perform Z-scaling, one may need to check if there is any skewness in the data and try to mitigate it, as it might cause problems during the model-building phase.
- **Train/Test Split:** For validation, one can use the k-fold cross-validation method.
- **Model-Building/Hyperparameter Tuning:** This is the final step at which different models will be tried and fine-tune their hyperparameters until desired level of performance is achieved.

Before building the model Class imbalance in the data needs to be handled. This can be done using SMOTE or ADASYN which are the most popular ways to handle class imbalance and see which gives better result.

Model Selection algorithms-

There are various models to choose from like logistic regression, KNN, Decision trees, XGBoost.
Trying out these models the final one to be selected would be the one which predicts the fraud correctly based on Model evaluation Metrics.
While Model building focus will also be on hyper parameter tuning to get the best model.

Model Evaluation-

Accuracy may not always be 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. The model which gives best AUC ROC curve should be used. For ROC curve, the best threshold would be one at which the TPR is high and FPR is low which would mean the misclassifications are low.