Solution Approach

1. Data understanding and exploring

2. Data cleaning

- Handling missing values
- Outliers treatment

3. Exploratory data analysis

- Univariant analysis of all The Numerical Variables
- Distribution of the Numerical Variables
- Categorical Variables
- Cardinality of Categorical Variables

4. Prepare the data for modelling

- Check the skewness of the data and mitigate it for fair analysis
- Handling data imbalance as we see only 0.172% records are the fraud transactions

5. Split the data into train and test set

- Scale the data (normalization)

6. Model building

- Train the model with various algorithm such as Logistic regression, SVM, Decision Tree, Random forest, XGBoost etc.

7. Model evaluation

- As we see that the data is heavily imbalanced, Accuracy may not be the correct measure for this particular case
- We have to look for a balance between Precision and Recall over Accuracy
- We also have to find out the good ROC score with high TPR and low FPR in order to get the lower number of misclassifications