Question 2: **Predict Fraudulent Transactions**

Tanmoy Baidya, B.Tech Part-IV , IIT (BHU) Varanasi

Methodology:

1.Missing Values Imputing

None of the numerical variables are having missing values. So, there is no need to preprocess the numerical variables. For categorical variables "back-fill" method is used to impute categorical missing values on both training and test data.

2.Label Encoding for Categorical Values

For homogenous label encoding, an array (named as labels) is created usuing numpy union1d method, which contains all the unique labels present in training and test data.Then LabelEncoder is fitted to that array and all the categorical columns (Column 1 – Column 18) are transformed.

3. Modelling:

In modelling stage, first feature scaling (Standardization) is done and then tree based model Random Forest and XGBoost is used. And for final submission 60% weight of XGBoost output and 40% weight of RandomForest output is taken.

4. Model Hyperparameters :

model_RF= Max Depth: 10,N Estimators: 100
model_XGB= Max Depth:10 , N Estimators : 150

5. Test Data Score : 0.73494 (AUC-ROC Score)

6. Tools Used:

Python 3.6 (Anaconda Distribution), Jupyter Notebook 5.2 , Pandas 0.21 , Numpy 1.13, scikit -learn 0.19, seaborn 0.8, matplotlib 2.1, scipy 1.0, xgboost 0.6a