**About Dataset Given and Data Handling**

* The given dataset is of 33000 rows and 45 columns , from which the columns 25 to 33 columns have lot of missing values , So we dropped them and filled others with median
* Found the Outliers in data replaced them using Outlier capping
* Dropped the Features which has single values
* Encoded the categorical using Target Encoding Technique.
* Used Heatmap and VIF to find the Multicollinear features dropped them using VIF where VIF>5
* From the graph we can say that the target is imbalance
* From the Scatterplot we can say that there is no pattern from which we can separate the classes. So we can say it’s a non linear data

**Data Preprocessing**

* Used RobustScaler for scaling the features
* As the data is non linear and imbalanced . I used Tree based algorithms, Bagging and Boosting , instead of using SMOTE to balance the data. By which the duplicate data will be created using KNN which has no use in model training

**Model Selection**

* Trained the different Tree models , Boosting models , Bagging and SVM with the “RBF” kernel.
* Selected CatBoostClassifier as the final Model because it has nice recall compared to other classes. As in recall is the accuracy of individual classes
* Selected top 15 feature using SelectKBest as we don’t have the Feature Details
* Tuned the model using GridSearchCV with selected features and trained the model.
* Used the same model for Predicting the Test csv.
* For Train csv we got the AUC of 0.77. Which is acceptable without doing much feature engineering and having domain knowledge to transform features.