Project Report

Data Understanding:

In the given train dataset there are total 11 feature columns and 245725 observations and in test dataset there are total 10 cols and 105312 observations. Most of them are don't have any missing values. And also most features are categorical in nature.

There are presence of missing values in following columns:

=>Credit_Product - 41847

Missing value Imputation:

As I mentioned above 1 column Credit_Product having "Yes" and "No" categorical values. I tried to change null values with both "Yes" and "No" values but it didn't increase auc_score, so introduce new category type as 'Unknown'.

Data Preprocessing:

After completing data cleaning part on the given dataset. In Avg_Account_Balance values are left skewed so I use log transformation to convert it into normal distribution And then encoding of the categorical feature using label encoder for high cardinal feature column and mapping function for binary feature columns.

Model Building Approach:

I tried to build 3 different model [lightgbm, catboost] without any feature engineering. Since it is classification problem having most of the features are of categorical and we can easily handle the categorical columns. I use K-fold method to prevent overfitting.

ROC_AUC_SCORE on different models:

1. Lightgbm - 0.8730
2. Catboost - 0.8722

I select lightgbm as my final model because it increase roc_score slightly.
.