Approach: Analytics Vidya - Machine Learning Summer Training Hackathon 2022

by Purushottam Behera (purushottambehera44@gmail.com)

Prediction of Loan Default:

Data Analysis:

First, we check the data type of the dataset's features to know whether they are **categorical** or ***numerical data.***

The data is imbalanced; the number of customers who default is significantly smaller than the number of customers who do not.

Then I find that only the education column has null values.

Data Preprocessing:

Missing values in the education column are filled using the MissingIndicator along with IterativeImputer and CatBoostRegressor.

GaussianMixture was applied with n_components = 3 in ‘age’ and ‘no_of_columns’ features and n_components = 2 in ‘no_of_curr_loans’.

Feature Engineering:

I added two new features to the given dataset.

1. diff_amount = difference between 'asset_cost’ and ‘loan_amount’
2. loan_complected = difference of ‘no_of_loans’ and ‘no_of_curr_loans’

Model Training:

I tried several base models, such as XGBClassifier, CatBoostClassifier, LGBMClassifier, and RandomForestClassifier, on the dataset. I passed the parameters (like class_weight='balanced') in the model to handle imbalanced data. Among these models, RandomForestClassifier gives the best macro f1 score. Therefore, this model is used for final submission.