From team S.A.D 2.0

Govind Anjan

Dennis Sagayanathan

Danish Arshad

Approach for the given problem statement.

1)Model planning: because our data contains more than 25000 rows and more than 5000 columns with null values included, we need to take an approach that should be time efficient, and the model needs to be accurate enough, so we went with XGBClassifier since that is the only algorithm that we found to work pretty well in a large dataset. We have also tested a few other algorithms, but initially, the results that we got using XGBClassifier were the most promising.

2) The preparation of the data and the preliminary processing:

We start by inputting our train data, then we proceed through a data cleaning process, during which we change nan and null values to 0 and accept infinite values (np.inf and np.-inf) up to the maximum value that float32 can handle.

3) We divided the data into an 80:20 ratio of train data to test data since the amount of the real test data that was sent to us by the organiser was equal to the size of the 20% test data.

4) The construction of the model: in order to develop our model, we used optuna to attempt many parameters at once and establish a range of values. This allowed us to experiment with all of the different possibilities. We have utilised the roc\_auc score and compare tool in order to opt for a better approach or choose a higher score than what we are currently obtaining for that particular algorithm. This has allowed us to get a higher overall score.

5) Operationalize: After obtaining the result, we retrain the model using the whole dataset for the last time in order to predict the values from the supplied test dataset and output results along with the predict provability.

TOOLS USED:

* XGBClassifier (ML model)
* Optuna (For finding the best parameters)
* Sklearn (for finding all performance metrics)
* Pandas (for data cleaning and manipulation)
* Python (programming language used to solve the problem)
* Jupyter notebook

Note: For the problem situation most ml model might not work if we don’t deal with infinity and null values in the dataset. So it is mandatory to work on the dataset Before hand with working on model building.