Health Insurance Lead Prediction

A brief on the approach

For this problem I have used **Logistic Regression & Cat Boost technique** for predicting the ‘Probability of Customer showing interest (class 1)’.

I have used Cat boost primarily because it provided categorical feature support i.e., it does not require encoding of the categorical data.

Data pre-processing

First, I have imputed missing values of the only column ‘Credit\_Product’ with ‘Unknown’.

Then I have trained the data by splitting train data in 4:1 ratio.

I have provided a list of columns to Cat Boost Classifier for Categorical columns.

After getting the best roc\_auc\_score, I have trained another model without splitting the data to get the results.

Then I predicted the results and consolidated it with ‘ID’ column to get the solution file.

The final trained model looks like this

