Approach Document for Job-a-than May 2021.

The following document describes the approach taken for the solution submitted in Job-a-than May 2021.

1. Problem Statement

Predict the probabilities of a Happy Banks' customer being potential lead for their credit cards.

2. Datasets

There were 2 datasets given such as train and test datasets. Train dataset had 11 Columns and test dataset had 10 Columns (Except Is_Lead Column). Following are the key highlights.

- There are 2,45,725 records in Train and 1,05,312 records in Test files.
- The datatypes of all the columns are mathing to their respective values inside. No need to modify the data types.

> Datasets are imbalanced.

3. Data Preprocessing Results

As part of Data Preprocessing which includes data cleaning, following assumptions were made.

- Column ID is the only column with all single values column and can be eliminated for ML modelling.
- Columns can be split based on the nature of the data inside as follows.
    - Categorial Columns - Gender, Region_Code, Occupation, Channel_Code, Credit_Product, Is_Active.
    - Numerical Columns - Age, Vintage, Avg_Account_Balance
    - Target Column - Is_Lead

3.1 Missing Value Treatment

Out of all the 10 Columns, it was only Credit_Product column had missng values. Following are the key highlights. * Expect Column Credit_Product, rest all the columns are not having any missing values in both Train & Test datasets. * There are missing values in Credit_Product which is significantly high (11.93% of entire column). * Because of the high volumn, it is not a good idea to drop all rows having missing values. * Choosing the right missing value will seriously affect the final prediction. * It is important to check the correlation of Credit_Product with other columns and replace the missing values based on the majority groups.

> It is highly dangarous to replace all the missing values with either 'Yes' or 'No' in Credit_Product.

3.2 Approach to replace the missing values

Based on the Correlation plot and Mosaic Plots (Check the code) following are the key inferences.

- Credit_Product mosaic chart indicates that the customers having credit products already had shown better interests (lead) to get credit card compaired to those who never had a credit product previously.
- Credit_Product has lightly better correlation with Channel_Code.
- Credit_Product has high correlation with target Column (Is_Lead). Once again Choosing right value for missing values will be significant.

Based on above above points, it was decided to take following 2 approaches for missing values.

- Approach-1: Replace the missing NaN with another category called Unknown. This is safest method compared to any kind of assumptions.
- Approach-2: Replace the missing NaN with marority groups. Here the missing value will be replaced with the mode value of Credit_Product among the values of groups (Credi_Product - Group by Channel_Code, Occupation and select the mode value).

4. ML modelling

As the dataset is imbalaced and involves many categorical values and Non-normal Numerical values, it is important to choose the ML model algorithm accordingly. Following 2 models have been choosen for further modelling.

1. Categorical Naive Bayes (CNB)
2. Light GBM (LGBM)

> Note: Both the models doesnot required any feature scaling.

After the experiments with ML modelling, following are the key highlights.

- Model CNB had good Recall and were able to predict more Positives cases.
- Model LGBM had a good Precision and were able to predict more Negative cases.

> The final result Is_Lead probability was the mean of above 2 models.

Both missing value approaches gave a predictions with accuracy more than 85% and ROC-AUC more than 0.8. Following are the key highlights

Approach-1

Following are the evaluation metrics for Approach-1 (Replace all the NaN with Unknown)

Model	Metric	Value

Model	Metric	Value
LGBM	Accuracy	0.8628751653270933
LGBM	Precision	0.8096653151678461
LGBM	Recall	0.5515886631896788
LGBM	ROC AUC	0.8811327582003137

Model	Metric	Value
CNB	Accuracy	0.8570149557432089
CNB	Precision	0.6946186305330935
CNB	Recall	0.7088594564919023
CNB	ROC AUC	0.8849188362974836

> Combined - ROC AUC 0.8932805814280729

Approach-2

Following are the evaluation metrics for Approach-1 (Replace all the NaN with Unknown)

Model	Metric	Value
LGBM	Accuracy	0.860274697324244
LGBM	ROC AUC	0.8735506751185225

Model	Metric	Value
CNB	Accuracy	0.8098443381829281
CNB	ROC AUC	0.8352426694443357

> Combined - ROC AUC 0.8623644970824206

Final Verdict

After comparing both the approaches, it was obvious to see Approach-1 has better reach over Approach-2. The final AUC Score for Approach-1 was 0.85