Analytics vidhya job a thon solution Approach

Problem Statement :

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products and other offerings.
Given the Customer details (gender, age, region etc.) and Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.) of existing customers of a Happy Customer Bank predict if any customer shows Interest towards a recommended credit card.

Evaluation Matrix:

roc_auc_score across all entries in the test set.

Reading the data :

It has 3 csv files Train,Test and Submit.The shape of Train data is (245725, 11) and that of test data is (105312, 10).

Feature Engineering and Preprocessing

1)Imputing null values

Only one column has null values,It has 11 to 12 percent of null values .To impute the missing values different methods like least ocuurred value, most occurred value, model based imputation and imputing with -1(or any random value not in column) were used. Imputing with -1 gave better results.

2)Exploratory data analysis

2.1) After analyzing the missing values and plotting a nested pie plot of missing column values and Output variables It is found that 85% of missing values are present only in lead = 1 though it is a minority class(24% of column and lead = 0 is 76%). Since most of the null values were present in lead = 1 a new feature is_null was created to predict the lead better.

2.2) After plotting the boxplot of the occupation vs vintage it is found that the salaried class has significantly more outliers than other classes. A feature is_salaried is created to better represent it.

2.3)After plotting the boxplot of channel_code vs vintage it is observed that X1 and X4 have very low Interquartile range to represent them two extra features is_X1 and is_X4 are created.

3)Feature binning

3.1) After printing the value count of the features Vintage and Age it is observed that many consecutive values have same number of count .To represent them properly and to make model much more robust these features are binned into multiples of 10(ex 0-9 as 0,10-19 as 1 ...90-99 as 9).

4)Polynomial features

4.1)New feature Balance/Vintage is created by dividing Avg_Account_Balance by Vintage.

4.2)New feature Balance*Vintage is created by multiplying Avg_Account_Balance and Vintage.

5)Label Encode all the categorical values

All the categorical values are label encoded as we are using tree based models.

6)Create features by combining two categorical columns

Two categorical columns are combined to get a new feature column and all combined features are label encoded.

7)Creating features by grouping and aggregating

To Create new features selected categorical columns are grouped and aggregating those those groups with remaining columns using specified functions.We get lots of features by this method and many of them are highly correlated. Features that have correlation greater than 0.9 are removed.

8)Frequency encoding categorical features

All the categorical columns are frequency encoded to give better feature representation.

9)Removing null valued columns if any

After creating aggregated features and many other types of features we may get null values in some columns it is better to remove such columns.

Building the Models

Here we are building two different models with the same sets of inputs.

Input features to both the models

Inputs to this model are features from exploratory data analysis, feature binning, polynomial features, combined categorical features, grouped and aggregated features, frequency encoded categorical features and numerical features.

Models

Xgb classifier and catboost classifier are trained with training data and are tuned with suitable hyper parameters. Respective test probabilities are predicted

Final Output

Both the test probabilities are multiplied with suitable values to get final results. Multiplication values are found using cross validation.

Iliyas H Indikar iliyasindikar@gmail.com