Credit Card Lead Prediction

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, among other offerings.

The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.

In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Now, the bank is looking for your help in identifying customers that could show higher intent towards a recommended credit card, given:

- Customer details (gender, age, region etc.)
- Details of his/her relationship with the bank (Channel_Code,Vintage, 'Avg_Asset_Value etc.)

Solution Approach

The solution approach is divided into 6 parts as follows:

1. Data importing and Data understanding
2. Data Cleaning
3. Exploratory data analysis
4. Data Preparation for Model Building
5. Model building
    a. Logistic Regression
    b. Decision Tree
    c. Random Forest
    d. XG Boost
    e. LightGBM
    f. Stacking
6. Model evaluation on Unseen (test file) data

1. Data importing and Data understanding

- As a first step, train data given as part of problem solution is imported as dataframe
- Next step was to understand the data by inspecting number of rows and number of columns present in the data.
- Also, which variable is of which data type and how each variable is distributed is checked.
- At last, checked how the response variable “Is_Lead” is distributed.

2. Data Cleaning

- As part of data cleaning activity, we first checked that are there any null values present for any columns
- We identified “Credit_Product” has more than 11% as null values. We imputed null values by ‘Unknown’ and retained all the data.
- After imputing, we get the cleaned data set for further analysis

3. Exploratory Data Analysis

- For Gender, Region_Code, Occupation, Channel_Code, Vintage, Credit_Product, Is_Active columns
    - We did univariate analysis by plotting countplot
    - We did bi-variate analysis by taking “Is_Lead” into account
- For Age and Avg_Account_Balance columns
    - We did univariate analysis by plotting histogram and boxplot

4. Data Preparation for Model Building

- For Gender, Region_Code, Occupation, Channel_Code, Credit_Product, Is_Active columns we performed one hot encoding by creating dummy variables
- We also removed ID column that is not required for Model building
- Then, we checked correlation between these variables to check if any variables are highly correlated with each other. No two variables have more than 0.70 correlation value.
- To evaluate the model on unseen data, we partitioned the given data into train and test in the ratio of 70:30

5. Model building

- We build five different models
    - Logistic Regression
    - Decision Tree
    - Random Forest
    - XG Boost
    - LightGBM
    - Stacking
- After building models we performed evaluation on partitioned test dataset using “roc_auc_score”

- We then compared the roc_auc_score and we identified **“LightGBM Classification”** model was giving better result, and we selected it as our final model

6. Model Evaluation on unseen data

- As part of evaluation on unseen data, we imported the test data provided as part of problem statement.
- We then performed data cleaning activities as done on train data.
- We prepared the data for model building by creating dummy variables for categorical variable using one hot encoding.
- And finally, we predicted the probabilities of the response for the customer