APPROACH USED:

Submitted By: Aruna Durai

Email ID – aruna.durai@outlook.com

I used Logistic regression model to predict if a customer is potential lead or not.

Steps followed:

1. Data Reading

2. Data Understanding: This step involves routine data check:
    - Inspecting the columns, shape, and datatypes of variables
    - lead.info () - list down all the columns along with name, no of non-null values, datatype, memory usage
    - lead. describe ()- describe dataset's mean, std, min,25%,50%,75%, max.

3. Data Cleaning

    The following steps will be done in data cleaning process:

    - Calculating missing value percentage for each column
    - Dropping the columns with high percentage of missing values.
    - Checking the unique category for each column.
    - If the columns are highly skewed with one category, such columns will be dropped. Combining different categories of the columns with less percentage values into "Others" category.
    - Imputing the column with least missing values percentage.
    - Checking for duplicate rows and dropping them if they are present.
    - Converting to correct datatype
    - Finally Checking for the number of rows retained after performing all the above steps.

    Result:

    - Only Credit_Product column had missing values. As it is a categorical column, I imputed it with Mode(‘No’).
    - Duplicate rows are not present.
    - Occupation column had ‘Self_Employed’, Entrepreneur','Self_Emp',’Others’. Entrepreneur' is also self-employed. So, combined Entrepreneur','Self_Emp' to Self_Employed
    - For Regional_Code, I combined categories with very low row percentages (<2%) as it does not make sense to create dummies for such rows.
    - Converted Is_Lead to category type.

4. Data Visualization -EDA

   a) Univariate Analysis of numerical columns:
      - The ‘Avg_Account_Balance’ is skewed.
      - Age and vintage don’t have outliers.

   b) Bivariate Analysis of Numerical columns:

      - The median of age of potential leads are around 50 years while 40 years for non-potential leads.
      - In vintage, the converted leads are in higher proportion in Q1 to median range that is roughly 25 to 70 months vintage period while in non leads, the high proportion of people are in 30 to 60 months vintage.

   c) Univariate Analysis of Categorical columns:

      - Data have more male than female. Many people have salaried occupation. Highest population is in region code 268.Channel code X1 have highest proportion

   d) Bivariate Analysis of Categorical columns:

      - The Male proportion is more in both leads and non leads class.
      - Salary people are less likely to turn into leads. So, it is necessary to focus on this class of grp.
      - X3 channel_code have highest leads and proportion of leads to non leads to less. Whereas in X1, the proportion of leads to non leads are high. So, bank need to come with actions keeping the above points into consideration.

   e) Multivariate Analysis of Categorical columns: None of the columns are very strongly correlated. So, no need to drop any columns.

5. Outlier treatment: Now from above EDA 5a , I can see that ‘Avg_Account_Balance’ is skewed. Transforming Avg_Account_Balance column to Normally distributed column.

   Method used: Power Transformation. I can try with log transformation also in future.

6. Data preparation before modelling:

   The following steps will be done as part of data preparation.

   - Converting binary variable (Yes/No) for columns Credit_Product','Is_Active’ to 1/0
   - Mapped Male with 1 and Female with 0 in Gender column
   - Create dummies for categorical columns.
   - Perform Scaling- I used StandardScaler method. I could have use min max scaling also as there are no outliers.

7. Mode Building:

- I used logistic regression model.
- After creating dummies, I was left with 30 columns.
- I used RFE was choosing 15 columns. I checked p value and VIF also.
- Our final model has p value and VIF less than 5%, which indicates that the variables I chose are significant, not correlated and thus, model is stable.
  (Note: I want to try using PCA also and try different algorithms. Because of lack of time, I have gone with basic model)

My model is quite stable. I got ROC value as 0.73 and accuracy, specificity, and sensitivity as 66% (cut-off probability is 0.275)

I am well sure that I can improve the performance if I try different techniques.

MODEL SUMMARY:

From our model, we can conclude following points:

Below are the variables used in model to predict the lead:

- Channel_Code_X3, Channel_Code_X2, Channel_Code_X4
- Age, Vintage, Credit_Product
- Occupation_Self_Employed, Occupation_Salaried,
- Region_Code_RG283, Region_Code_RG254, Region_Code_RG284, Region_Code_RG277, Region_Code_RG273, Region_Code_RG252, Region_Code_RG279

From EDA and model variables, it can be concluded that the bank should focus on more on

- Channel_Code_X3, Channel_Code_X2, Channel_Code_X4 because the proportion between lead and non-lead is less compared to Channel_Code_X1.
- Age feature also play important role. The median of age of potential leads are around 50 years whereas 40 years for non-potential leads. So, bank should focus on actions that attract non potential age leads also.
- In vintage, the converted leads are in higher proportion in Q1 to median range that is roughly 25 to 70 months vintage period whereas in non leads, the high proportion of people are in 30 to 60 months vintage.
- The proportion between lead and not lead is less for customer with active credit product even though the leads are higher if they do not have active credit product.
- Occupation with Salaried category have high proportion between lead and non-lead and self-employed have less comparatively and from above graphs we can see that, the self-employed category has higher leads.
- Region_Code_RG283, Region_Code_RG254, Region_Code_RG284, Region_Code_RG268 have high leads even though proportion to leads and non leads have high

- Region_Code_RG252 have least leads and remaining region model variables have low proportion between leads and non leads even though leads are less.

BUSINESS IDEAS:

- ❖ Age around 50 years, self-employed with Channel code X3, X4, X2 with vintage period around 2 to 6 years with inactive credit product(no) are potential leads.
- ❖ The bank should come up with new schemes to convert people into leads with age around 40 years, salaried category with vintage period 2.5 to 5 years with active credit product.
- ❖ People in all the region codes need to be focused. Either proportion of leads to non leads is high and leads are also high or proportion of leads to non leads is low, and leads are also low.