Approach Document

Credit Card Lead Prediction

By

Samsensurya S

Contents:

I. Exploratory Data Analysis:

- Identification of datatypes of different variables present in the dataset.

```
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   ID                   245725 non-null  object
 1   Gender               245725 non-null  object
 2   Age                  245725 non-null  int64
 3   Region_Code          245725 non-null  object
 4   Occupation           245725 non-null  object
 5   Channel_Code         245725 non-null  object
 6   Vintage              245725 non-null  int64
 7   Credit_Product       245725 non-null  object
 8   Avg_Account_Balance  245725 non-null  int64
 9   Is_Active            245725 non-null  object
 10  Is_Lead              245725 non-null  int64
dtypes: int64(4), object(7)
```

- To understand the distribution of data barplots and countplots were used.
- While looking at data distribution it was evident that there are missing values in Credit_Product column.
- In each category there were certain values that has significant influence on the dependent variable.
- Average account balance column had some abnormal values which were identified by boxplot
- There were no duplicate data since there was Unique ID which has high cardinality
- Out of 11 variables ID was the only variable which has no significant impact on the data.
- Channel Codes X2 and X3 has significant impact on data than X1 and X4
- Majority of the customer where distributed across five regions.

- Customers above 40 were highly likely to buy credit cards than customers below 40.
- Age and Vintage had correlation of about 0.6.
- Among occupation Entrepreneur had very less contribution than the other occupation

II. Data Cleaning:

- Credit_Product had missing values and when it was replaced by mode of the variable, since it's a categorical variable , it had very little effect since the mode is no and data became imbalanced because of higher percentage of No values.
- Missing values were replaced by value 'Unknown'. Hence forming three value variable.
- Average balance had extreme values. In order to remove the extremities Log function was used to remove the outliers.
- All the object type variables are type casted to Categorical variable and Int, Float to Numeric variable.
- All the above were parallelly done to testdata as well

III. Feature Engineering:

- Label Encoding is used for Gender, Is_active and Region code
- Region Code had high number of regions hence label encoder is used instead of One hot encoder
- One Hot encoding is used for Occupation, Channel code and Credit product and age.
- Train Data is split into 70% training set and 30% test set

- Training set data is scaled using Minmax scaler.
- All the above were parallelly done to Test data as well, wherein testdata is not split into training and test set.

IV. Modelling:

- Logistic Regression was used as machine learning model to train the data. No hyperparameter tuning was performed
- The Value of ROC_AUC_Score is 0.84.

Experimenting with other models such as KNN, Random forest , SVM.