ANALYTICS VIDHYALAYA : Job-A-Thon

Problem Statement : Credit Card Lead Prediction

1. Brief Approach

The problem statement provided us with set of training and testing data to classifying the interest of the customers towards Credit Card. To begin with, the preliminary step for training any model is data pre-processing. In this step, we handle values like duplicate values, missing values, null values, etc. We convert the categorical data into meaning full data using various methods like One-hot Encoding or Label Encoding. Post this step, our next step is Feature Selection i.e., selecting features that are necessary for training our model since all the features are not mandatory for training the data. For this step, we need to perform exploratory analysis of the data. We can do this in Python using various libraries like seaborn, matplotlib, etc. After extracting all the necessary features, we will continue with training our model. There are various algorithms which we can pick, but I opted for Decision Tree. My reason for opting for this algorithm are-

a) Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
b) The logic behind the decision tree can be easily understood because it shows a tree-like structure.

I also tried algorithms like Logistic Regression, k-NN, Random Forest but Decision Tree suits the best among all for the given training dataset.

Before training the data, I split my dataset into training and validation dataset with ratio 75:25. Once the training of my model was completed, I validated my data by calculating accuracy of the model and confusion matrix. Also, I tested my model against my validation dataset. Once completed I tested my model against the testing data and stored the result in a .csv file.

2. Data Pre-processing

a) To start with I checked the structure and datatype of each of the column –

```
ID                     object
Gender                 object
Age                     int64
Region_Code            object
Occupation             object
Channel_Code           object
Vintage                 int64
Credit_Product         object
Avg_Account_Balance     int64
Is_Active              object
Is_Lead                 int64
dtype: object
```

b) Describe the structure –

	Age	Vintage	Avg_Account_Balance	Is_Lead
count	245725.000000	245725.000000	2.457250e+05	245725.000000
mean	43.856307	46.959141	1.128403e+06	0.237208
std	14.828672	32.353136	8.529364e+05	0.425372
min	23.000000	7.000000	2.079000e+04	0.000000
25%	30.000000	20.000000	6.043100e+05	0.000000
50%	43.000000	32.000000	8.946010e+05	0.000000
75%	54.000000	73.000000	1.366666e+06	0.000000
max	85.000000	135.000000	1.035201e+07	1.000000

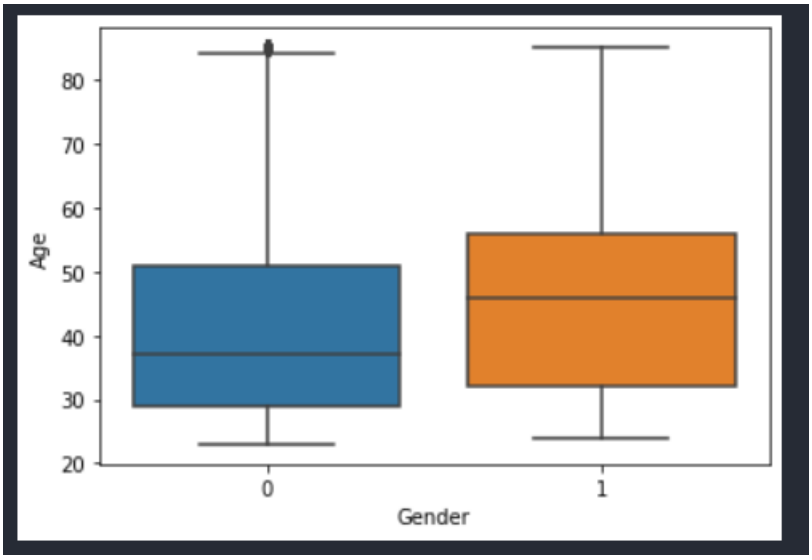
c) Next steps were to calculate the distinct values present in columns like Occupation, Channel_Code, Vintage, etc.
d) Check the null/missing/duplicate values in the data. As per analysis we had null values for Credict_Predict –

```
ID                        0
Gender                    0
Age                       0
Region_Code               0
Occupation                0
Channel_Code              0
Vintage                   0
Credit_Product        29325
Avg_Account_Balance       0
Is_Active                 0
Is_Lead                   0
dtype: int64
```

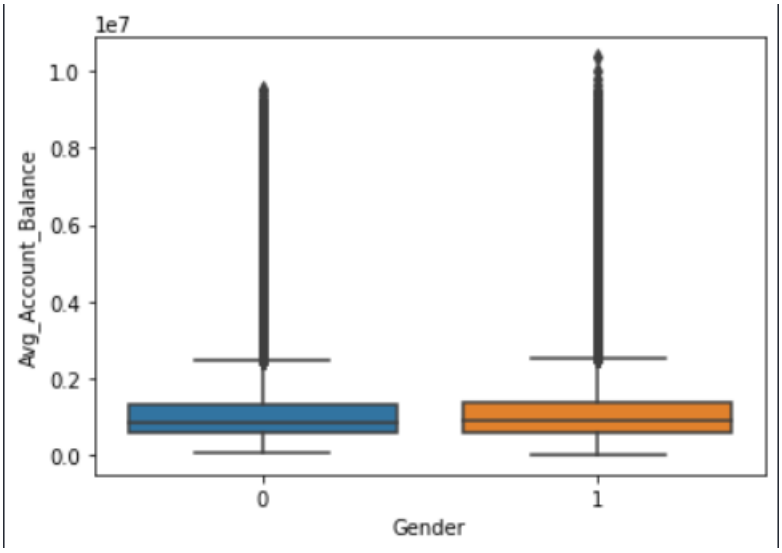
e) Replace the null value based the most frequent word depending on Is_Lead, Credict_Predict, etc.
f) Next step is to handle categorical data. I opted for Label Encoding for features like – Gender, Is_Active, Credict_Product and One-Hot Encoding for – Occupation, Channel_Code.

3. Exploratory Analysis of Data

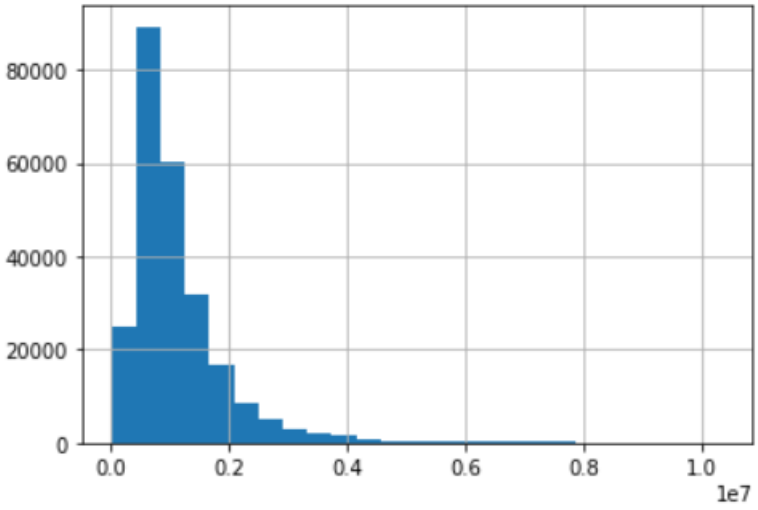
a) Relation – Gender vs Age



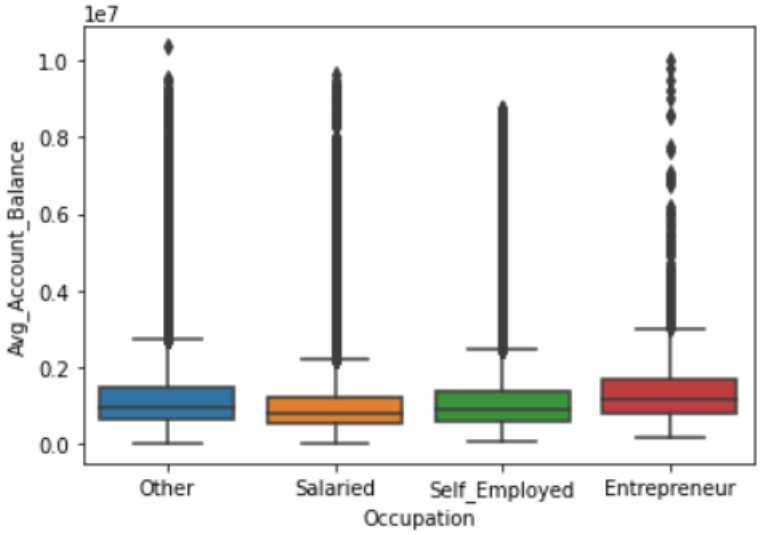
b) Relation – Gender vs Average Balance



c) Average Salary distribution



d) Relation – Occupation vs Average Salary



4. Feature Selection

Based on the preprocessing step and exploratory analysis of data, the important features for training are as follows –

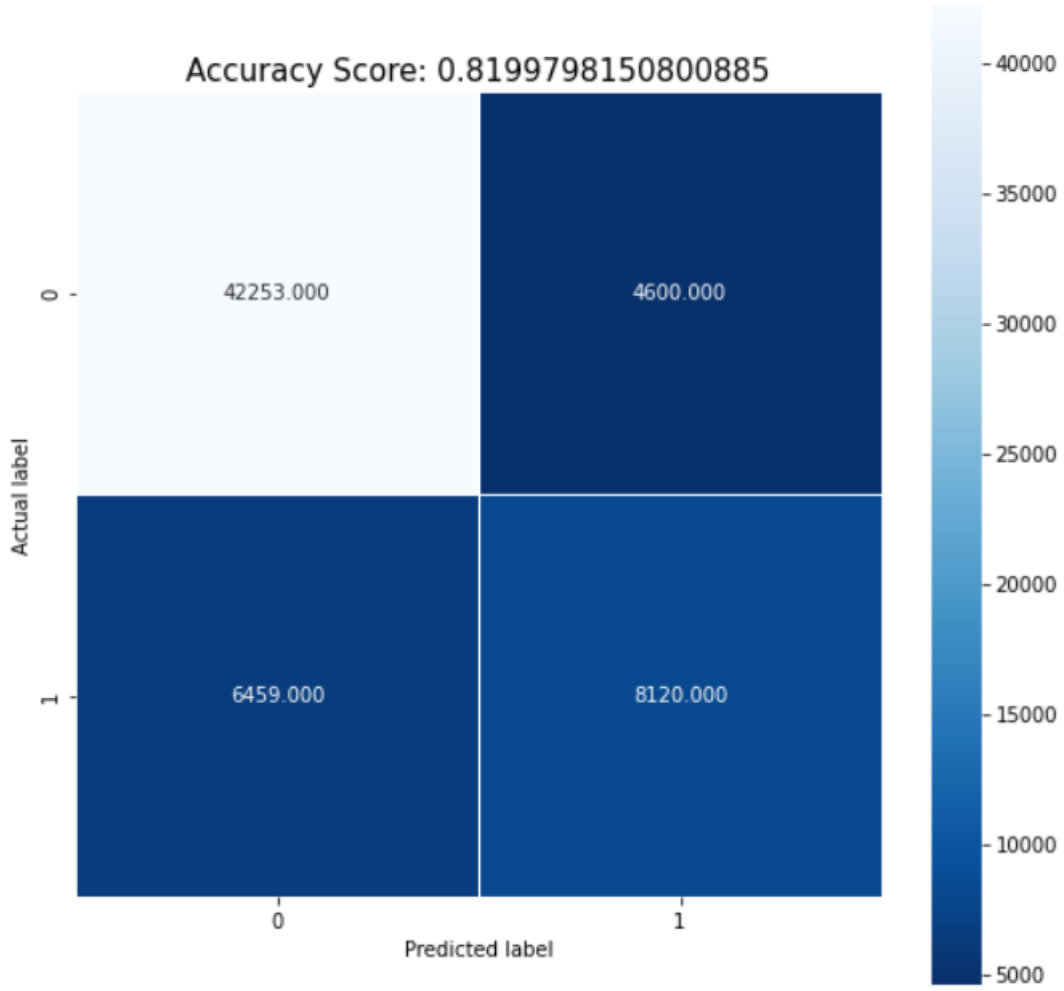
a) ***Gender –*** To observe the behaviour of both genders for showing interest in Credit Card
b) ***Credit_Product*** – This represents whether any customer already has a credit product or not. This also plays an important role
c) ***Vintage –*** Vintage for the Customer
d) ***Avg_Account_Balance*** – This feature is important as balance in the account related towards the interest for Credit Card
e) ***Is_Active*** – This denote if customer is active user or not. Active users have higher tendency of showing interest for credit cards
f) ***Entrepreneur, Other, Salaried, Self_Employed*** – Occupation of customers
g) ***X1, X2, X3, X4*** – Channel code of customers

5. Training Model

This is one the important steps. Selecting correct attributes for training the model is very important. Incorrect attribute selection can lead to underfitting or overfitting of the data. For decision trees, appropriate value for entropy and information gain is most important along with Gini value. After training the model, next step is to calculate the accuracy of the model. Below is the accuracy for all the different algorithms I used for training my model –

	Name of Model	Accuracy Score of Model
0	Logistic Regression	0.762681
1	k-NN	0.733657
2	Random Forest	0.790630
3	SVM	0.762681
4	Decision Tree	0.819980

To continue, calculating value of confusion matrix is necessary as it will further help us to calculate different values like Recall, Precision, etc. For my model, confusion matrix looks like –



6. Testing Model against testing data

As the final step, I tested my model against the given testing data and stored the predicted value in *submission.csv* file

Thank You