Credit Card Lead Prediction

Approach Document

<u>Prepared by : Abhishek Chowdhury</u>

Contents

Exploratory Data Analysis (EDA)

1. Missing Values

In this step which columns had missing values was checked and the following was observed :

- In train dataset, **Credit_Product** there is **11.93 %** missing data points.
- In test dataset, **Credit_Product** there is **11.89 %** missing data points.

```
In [7]:  1 train.isnull().mean()*100

Out[7]: ID                      0.000000
        Gender                  0.000000
        Age                     0.000000
        Region_Code             0.000000
        Occupation              0.000000
        Channel_Code            0.000000
        Vintage                 0.000000
        Credit_Product         11.934073
        Avg_Account_Balance     0.000000
        Is_Active               0.000000
        Is_Lead                 0.000000
        dtype: float64
```

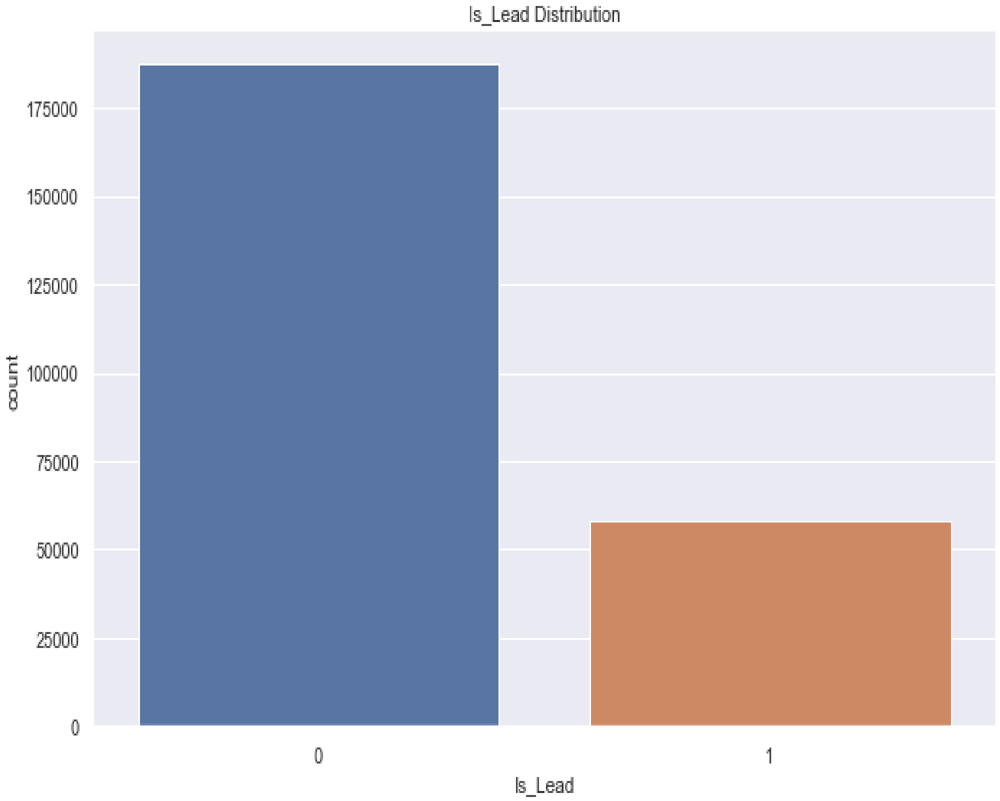
```
In [8]:  1 test.isnull().mean()*100

Out[8]: ID                      0.000000
        Gender                  0.000000
        Age                     0.000000
        Region_Code             0.000000
        Occupation              0.000000
        Channel_Code            0.000000
        Vintage                 0.000000
        Credit_Product         11.890383
        Avg_Account_Balance     0.000000
        Is_Active               0.000000
        dtype: float64
```

Exploratory Data Analysis (EDA)

2. Target Variable Distribution

The target variable **Is_Lead** has imbalance between the two classes 0 (76.27%) and 1 (23.72). We need to use over-sampling techniques like SMOTE to balance the dataset.



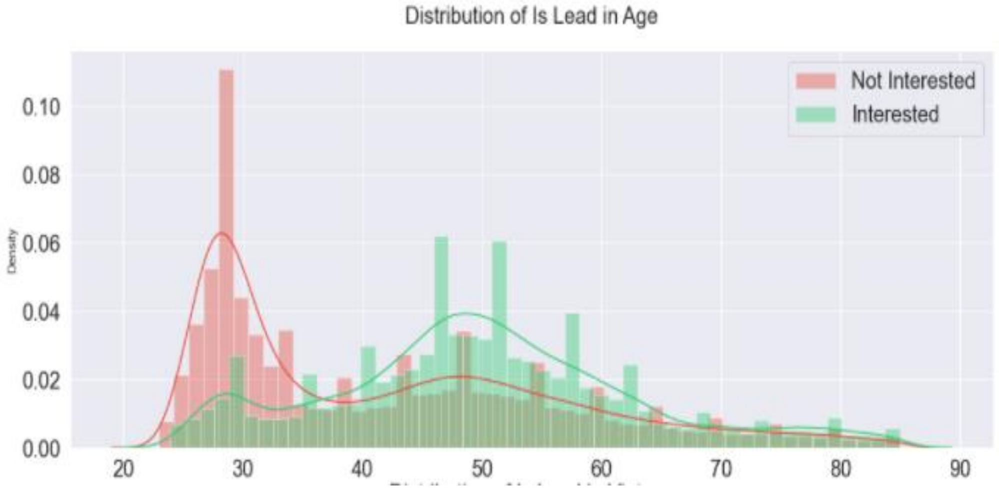
Exploratory Data Analysis (EDA)

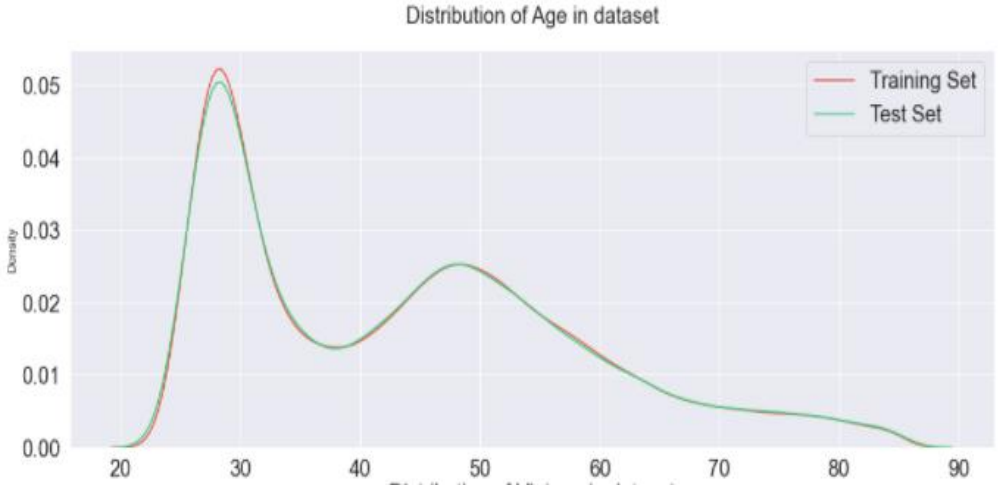
3. Numerical Features Distribution

In this step the distribution of the numerical features are checked and insights from the charts are inferred

3.1 Age

- The Age distribution in train dataset and test dataset is almost similar
- Customers aged between 40-60 have greater interest in credit cards.
- Customers in their 20s and 30s and less interested
- Age feature has a skewness of 0.619 and kurtosis of -0.441 in train set and a skewness of 0.628 and kurtosis of -0.423 in test set

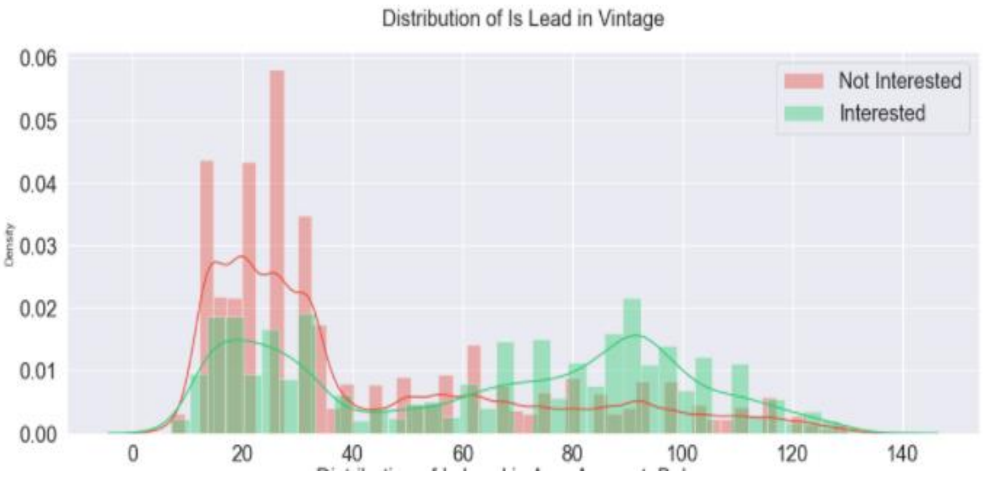


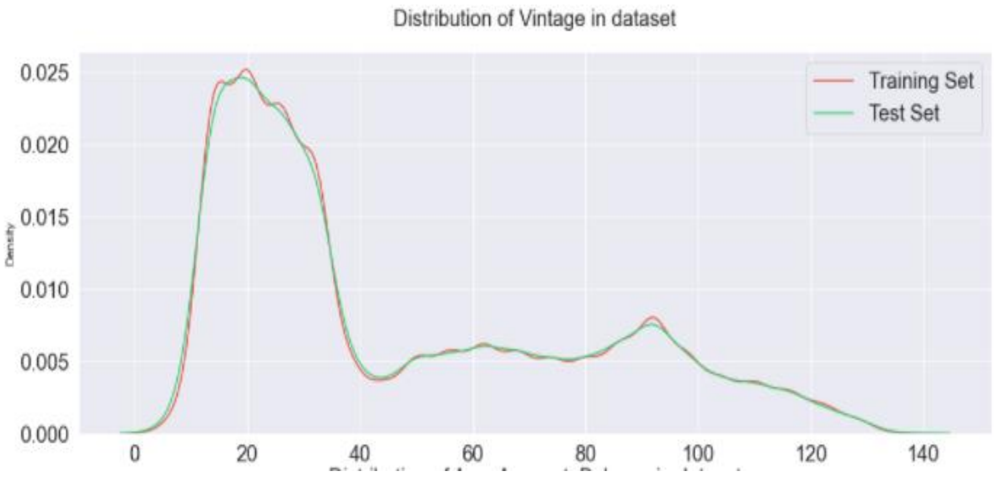


Exploratory Data Analysis (EDA)

3.2 Vintage

- Vintage feature has a skewness of 0.79 and kurtosis of -0.697 in train set and has a skewness of 0.791 and kurtosis of -0.689 in test set
- The distribution of the Vintage feature is very similar in the train dataset and test dataset
- Among the customer segment, who have accounts for a longer vintage period (80-100 months) are more interested to take up Credit Cards than their counterparts.
- Among the lower Vintage period customers (0-36 months) the proportion of customers not interested in taking up Credit Cards is more.



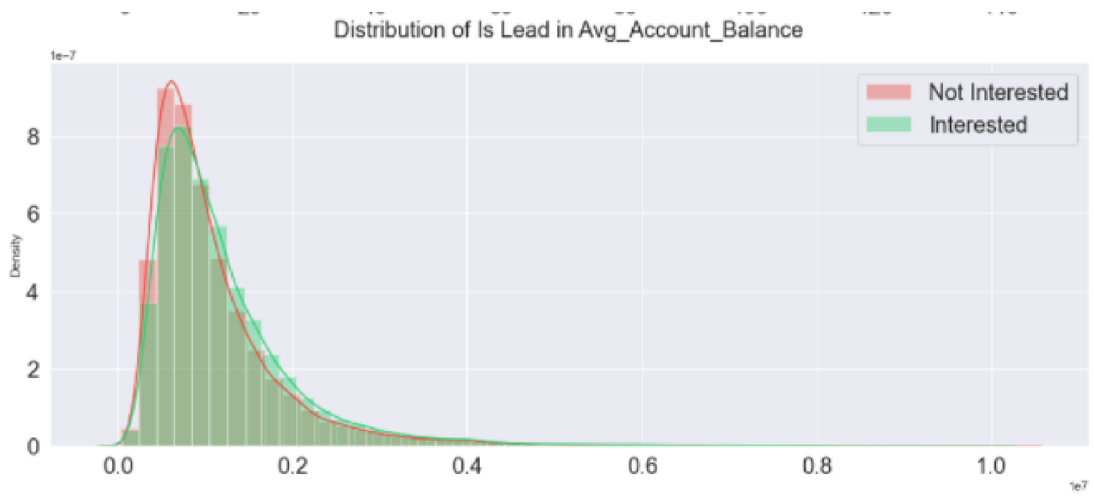


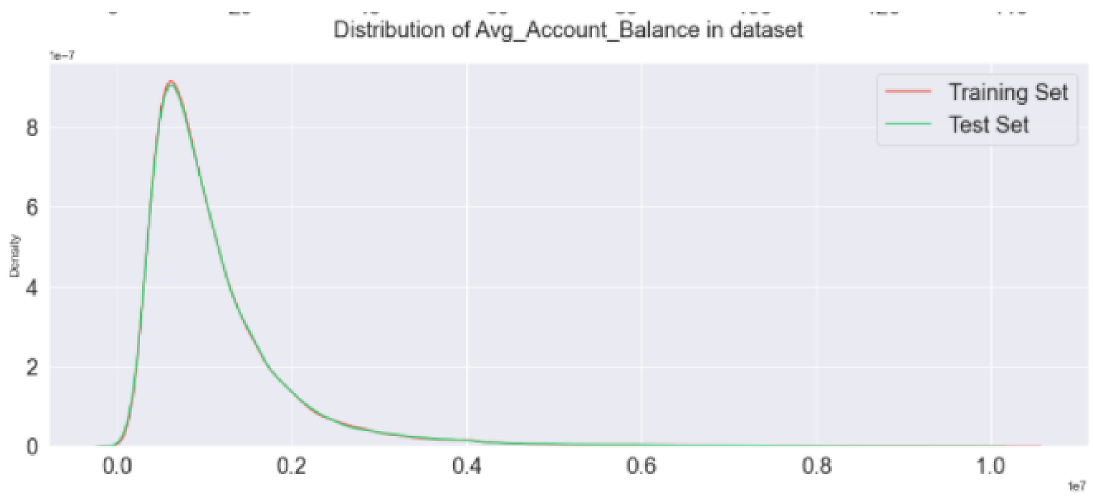
Exploratory Data Analysis (EDA)

3.3 Avg_Account_Balance

- The Avg_Account_Balance, Vintage distribution in train dataset and test dataset is almost similar.
- Avg_Account_Balance feature has a skewness of 2.969 and kurtosis of 14.305 in train setand has a

skewness of 2.998 and kurtosis of 14.43 in test set.





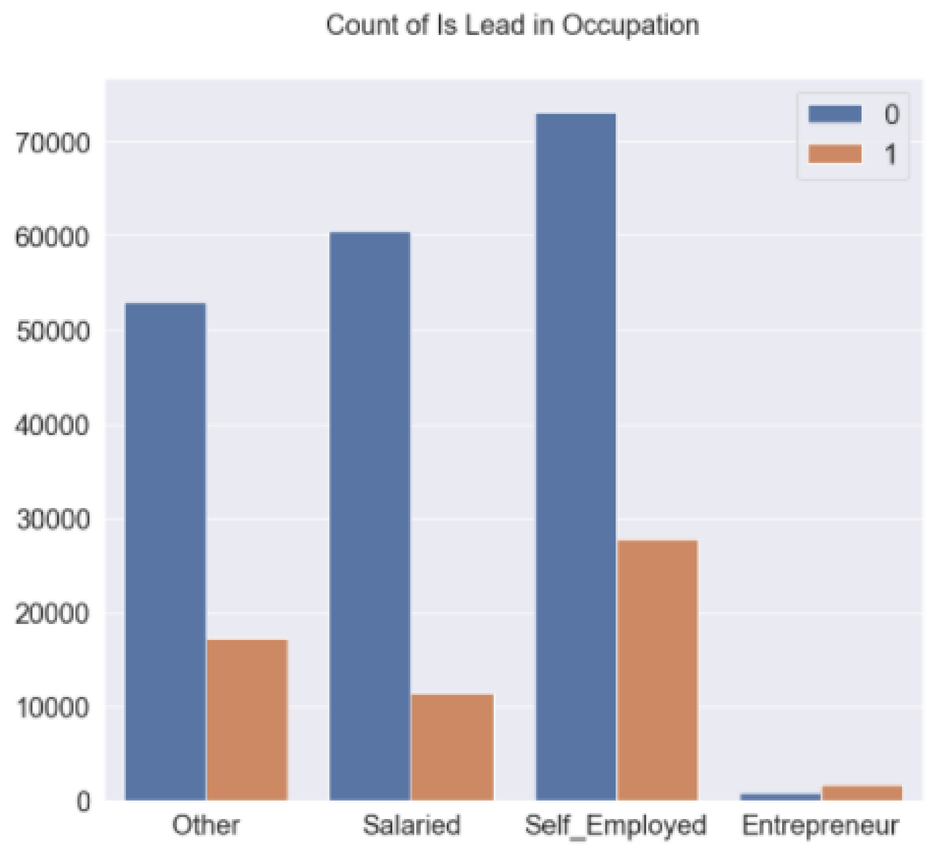
Exploratory Data Analysis (EDA)

4. Categorical Features Distribution

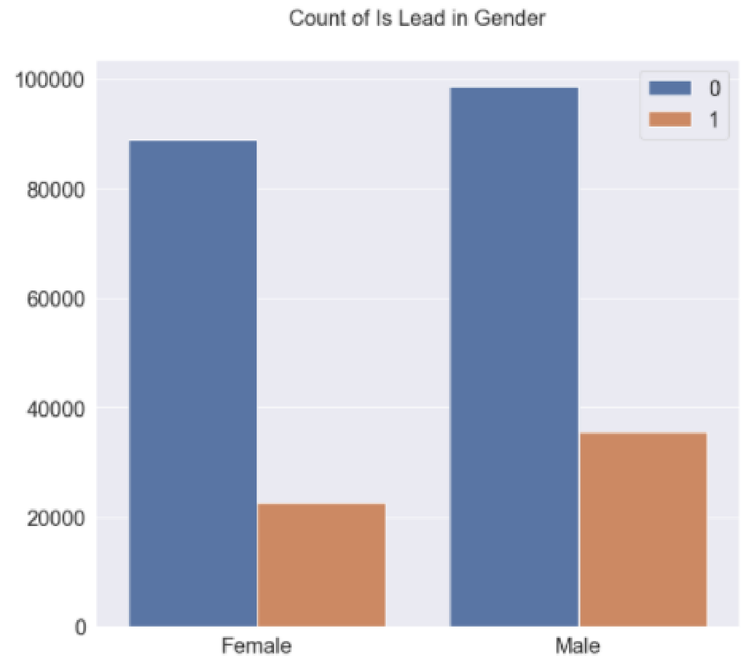
4.1 Occupation

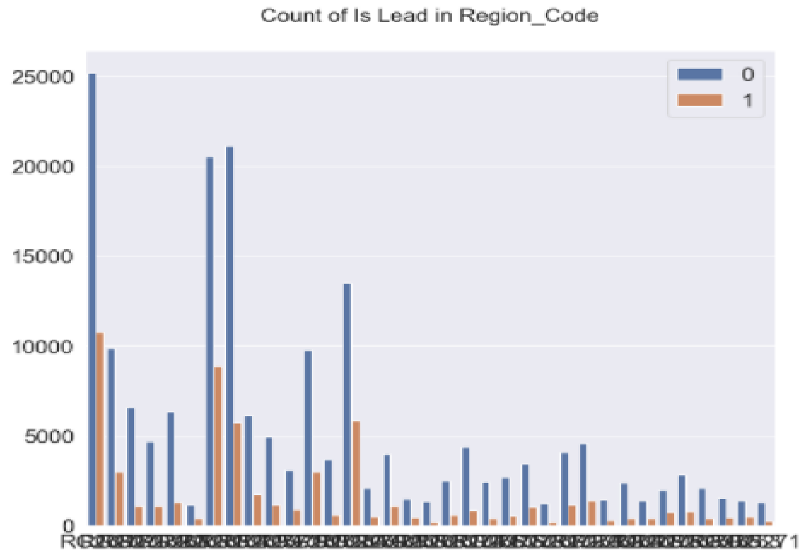
- Salaried person are less likely to take up credit cards. Only among Entrepreneur the number of customers interested to take up credit cards is more
- Only 2 Entrepreneurs don't have any credit product.
- 66% of total Customers falling in Entrepreneurial category in Occupation have shown interest in the past followed by 27.6% Self Employed, 24.5% in Others category and 16% Salaried.

• Age 40-65, salaried are interested to buy credit card. Most of the Entrepreneurs also seem interested but not all.



Exploratory Data Analysis (EDA)



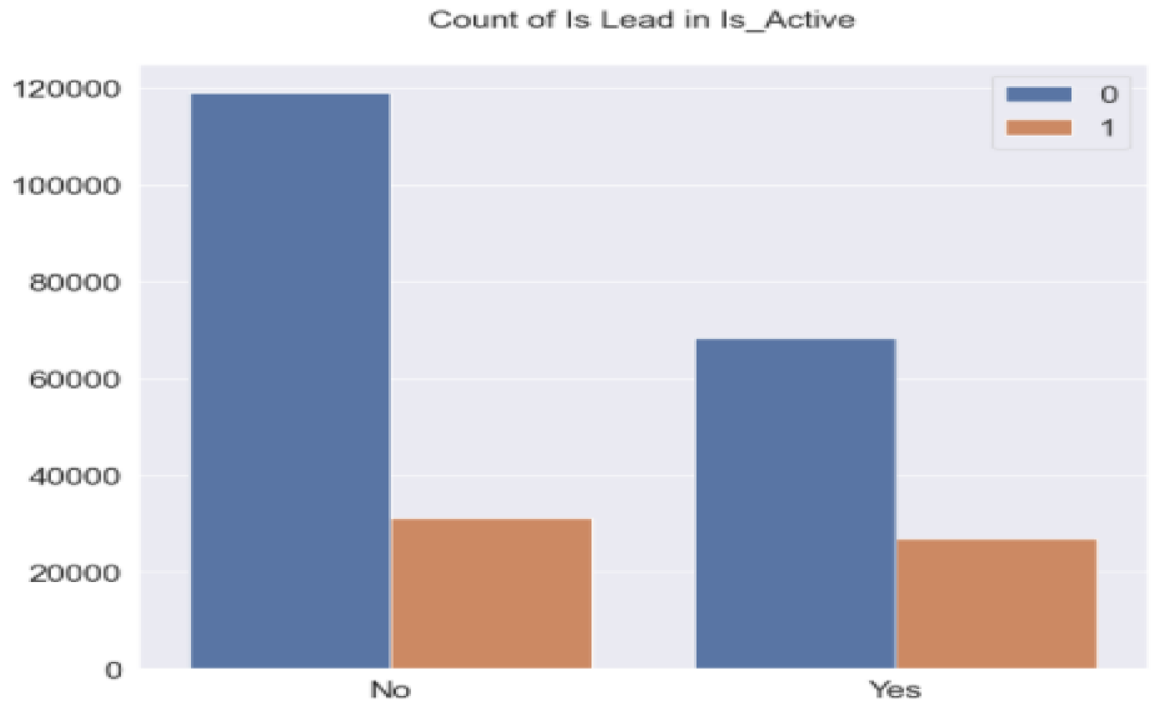


4.2 Gender

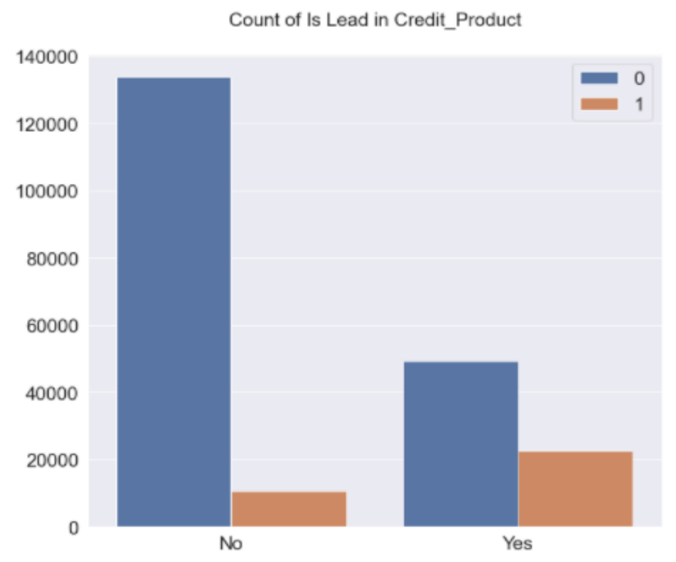
- Male customers are present more in the dataset than females

4.3 Is_Active

- Male customers are present more in the dataset than females

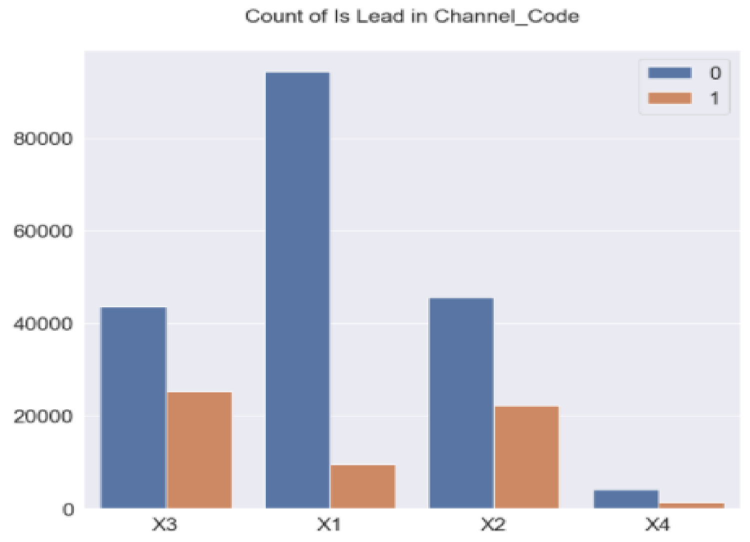


Exploratory Data Analysis (EDA)



4.4 Credit Product

- Number of Customers having credit products who are interested in Credit Card is more than those who donot have a Credit Product.
- Customers who already have any credit product are likely to buy credit card.



4.5 Channel Code

- Salaried people with Channel code X1 haven't shown much interest in the past.
- Customers from Channel X3 is the maximum group of customers interested in Credit Cards.

Exploratory Data Analysis (EDA)

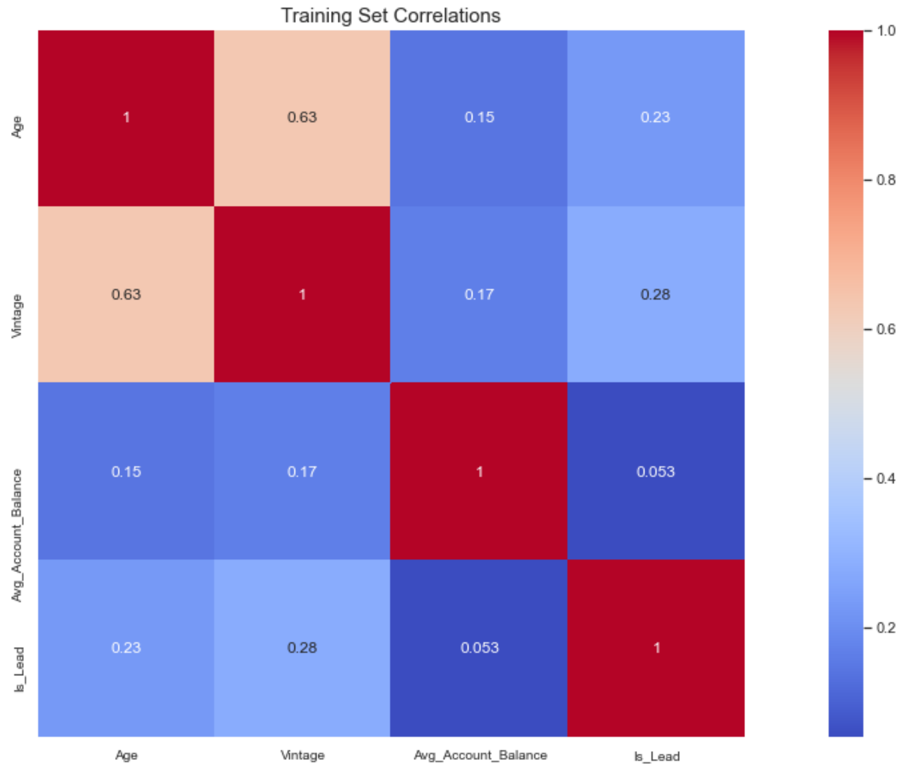
5. Outliers

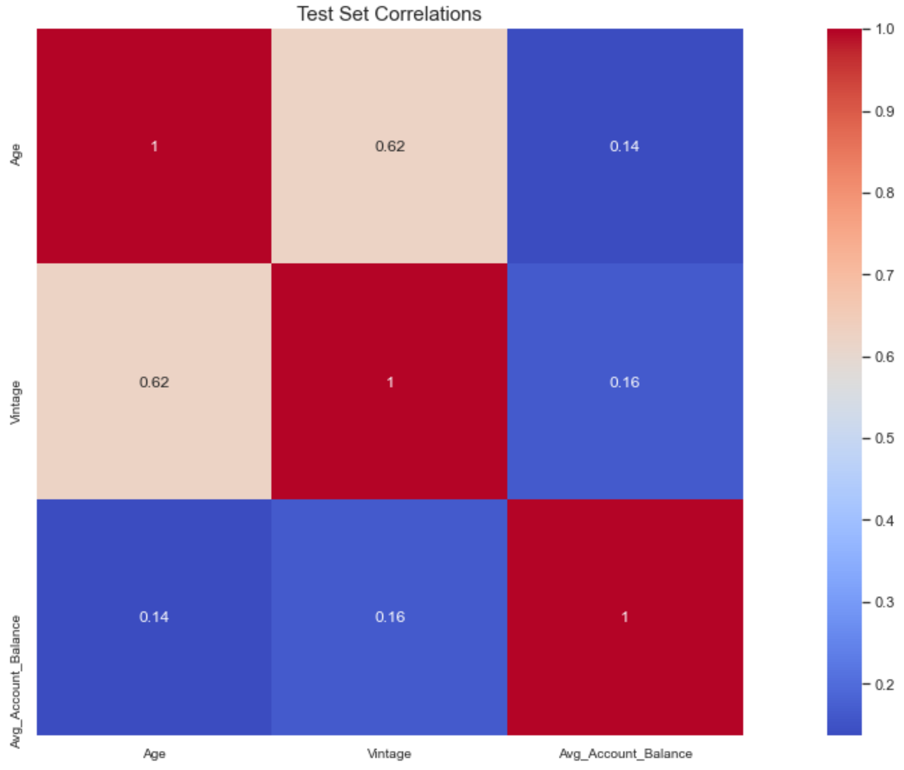
In the train and test dataset only ***Avg_Account_Balance*** has outliers. Other numerical features like Age and Vintage do not have any outliers.

Exploratory Data Analysis (EDA)

6. Correlations

Age and **Vintage** has highest correlation (**0.63**) both in train dataset and (**0.62**) in test dataset.



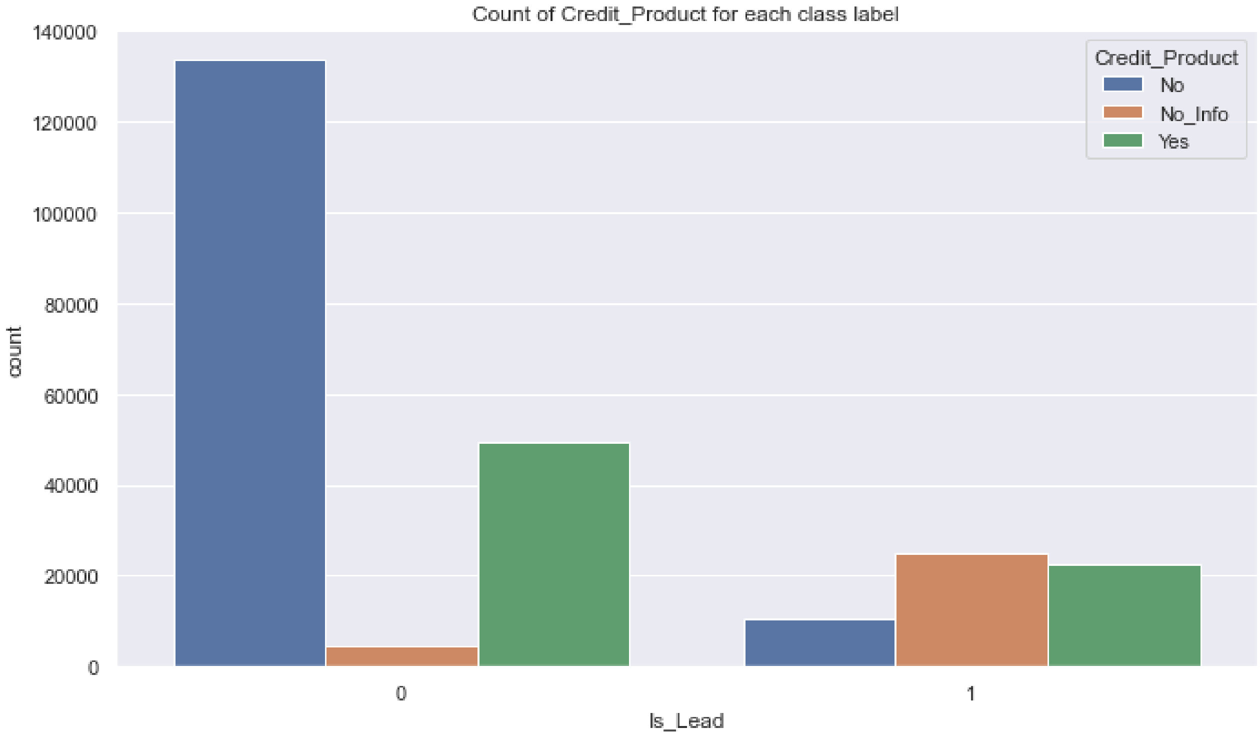


Data Cleaning

1. Handling Missing Values

The missing values in the feature ***Credit_Product*** feature both in train dataset and test dataset is imputed with ***No_Info***

*Distribution of **Credit_Product** after imputing **"No_Info"***



Feature Engineering

1. One Hot Encoding of Categorical Features

All the ***categorical columns*** were ***One Hot Encoded viz.*** *["Gender","Region_Code","Occupation", “Channel_Code”, “Credit_Product","Is_Active"]*

For One Hot Encoding ***pd.get_dummies() from Pandas*** was used and ***irrelevant columns like ID were removed***.

Oversampling (Handling Class Imbalanced)

There is a class imbalance observed in the target feature. About **76.27%** customers are not interested in credit card, and about **23.72%** are interested in credit card.

To address this issue Oversampling techniques like **SMOTE** is needed in order to balance the class imbalance.

After using SMOTE from imblearn the class imbalanced was removed using oversampling techniques

In [66]:

```
smote = SMOTE(random_state=42, n_jobs=-1)
X_train_smote, y_train_smote = smote.fit_resample(X,y)
X_train_smote.shape, y_train_smote.shape, y_train_smote.value_counts()
```

Out[66]:

```
((374874, 53),
 (374874, 1),
 Is_Lead
 0          187437
 1          187437
 dtype: int64)
```

Modelling and Hyperparameter Tuning

In this approach,

- For Modelling both Light GBM and XGBoost model is used.
- The dimensionality of data is low hence the tree based approach.
- For combining the predictions made by XGBoost and Light GBM, stacking is used
- The models were tuned by Randomized Search CV
- The predictions of the two models were ensembled using stacking

The XGBoost model gave a ROC-AUC score of 0.879 while the Light GBM model the ROC-AUC score was 0.876

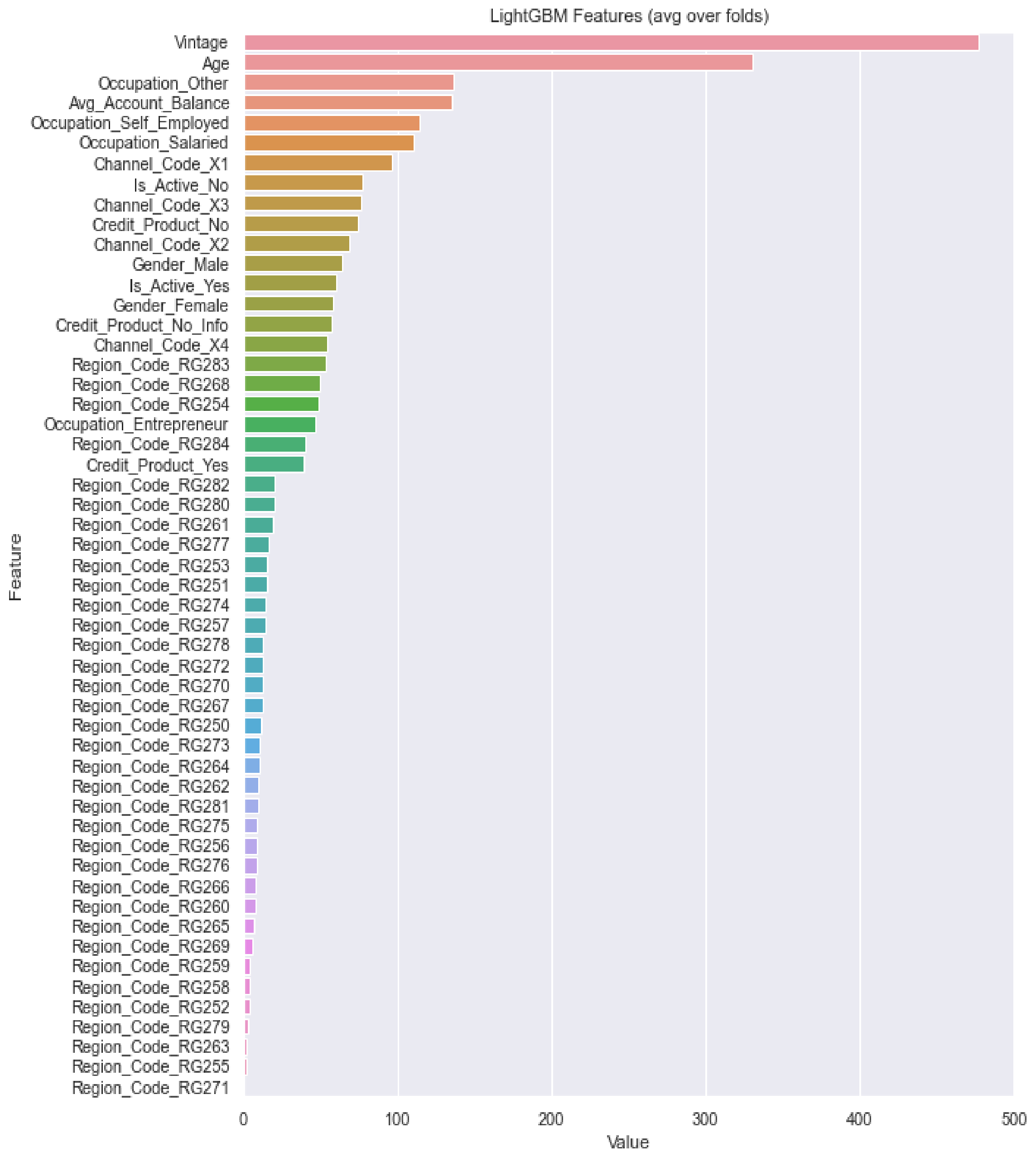
In this problem ROC-AUC score was used as evaluation metric.

Feature Importance

1. **Light Gradient Boosting Model**

For **Light GBM model** the top 5 features of importance were :

1. **Vintage**
2. **Age**
3. **Occupation_Other**
4. **Avg_Account_Balance**
5. **Occupation_Self_Employed**



Feature Importance

2. Xgboost Model

For the **Xgboost model** the top 5 features of importance are :

1. **Avg_Account_Balance**
2. **Vintage**
3. **Age**
4. **Is_Active_No,**
5. **Credit_Product_Yes**