HACKATHON APPROACH USED

CREDIT CARD LEAD PREDICTION

- *SOORYA P.G.*

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.)

Objective

In this project, we are supposed to build a machine learning model to solve the task at hand and then submit its solution along with the code file. To complete this project successfully, you are supposed to submit the solution with the EDA, feature engineering and model building part with a detailed explanation of your approach.

Data Description

Variable	Definition
ID	Unique Identifier for a row
Gender	Gender of the Customer
Age	Age of the Customer (in Years)
Region_Code	Code of the Region for the customers
Occupation	Occupation Type for the customer
Channel_Code	Acquisition Channel Code for the Customer (Encoded)
Vintage	Vintage for the Customer (In Months)
Credit_Product	If the Customer has any active credit product (Home loan, Personal loan, Credit Card etc.)
Avg_Account_Balance	Average Account Balance for the Customer in last 12 Months
Is_Active	If the Customer is Active in last 3 Months
Is_Lead(Target)	If the Customer is interested in the Credit Card 0: Customer is not interested 1: Customer is interested

Sample of the Dataset

The given dataset sample is shown in the table 2 where we are having 11 variables describing the customer details and other details based on their ID which identifies each customer. Based on these variables, we are going to find whether the customer will show intent towards a recommended credit card. We are also having some categorical variables like region code, Occupation, Gender, Channel code, Credit Product, Is Active and Is Lead which is the target variable. We are having around 6 categorical variables and 3 numerical variables.

ID	Gender	Age	Region _Code	Occupation	Channel_ Code	Vintage	Credit_ Product	Avg_Account _Balance	Is_ Active	Is_ Lead
NNVBBKZB	Female	73	RG268	Other	X3	43	No	1045696	No	0
IDD62UNG	Female	30	RG277	Salaried	X1	32	No	581988	No	0
HD3DSEMC	Female	56	RG268	Self_ Employed	X3	26	No	1484315	Yes	0
BF3NC7KV	Male	34	RG270	Salaried	X1	19	No	470454	No	0
TEASRWXV	Female	30	RG282	Salaried	X1	33	No	886787	No	0

The given dataset contains around 245725 rows about customers, and we have 11 columns to support each of those rows. Let us look at the basic understanding of each of the columns present in the dataset using the describe function, info function in python.

Basic Info of the Dataset

S.No	Column	Non-Null Count	Datatype
1	ID	245725 non-null	object
2	Gender	245725 non-null	object
3	Age	245725 non-null	int64
4	Region_Code	245725 non-null	object
5	Occupation	245725 non-null	object
6	Channel_Code	245725 non-null	object
7	Vintage	245725 non-null	int64
8	Credit_Product	216400 non-null	int64
9	Avg_Account_Balance	245725 non-null	object
10	Is_Active	245725 non-null	object
11	Is_Lead	245725 non-null	int64

From the above table 3, we can see the basic details about the columns of the dataset which contains information like the non-null values count, datatype of each variable, etc. We find that there are some missing values present in one of the variables present in the dataset.

So, in the process of EDA, we will be treating the missing values present in both the categorical Credit Product variables present in the dataset.

Exploratory Data Analysis (EDA) and Data Pre-Processing

We'll be doing EDA on the given dataset to find any insights which can be helpful for us to help the company in retaining its customers. We can check for missing values, duplicate values first for EDA before going for Univariate and Bivariate Analysis.

Checking for Missing Values and Duplicate Values

1. There is a variable called as ID which only serves the purpose of Unique identifier for each customer/account, and it is not needed for our EDA process as well as for the process of modelling. So, it is better for us to remove this variable. We can remove this variable in python using the drop function.

2. We can check for missing values present in the dataset using the is null function in python as shown in the result below. We find that one of the variables is having missing values. So, it is required for us to impute these missing values with either mean or mode based on whether they are categorical or numerical variable.

```
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
```

As we can see from above count, there are large number of missing values present in the dataset in only one variable. So, we will be treating those missing values with the help of mean/ mode depending on the variable being numerical or categorical.

But since, we have found some correlation between the target variable and the missing values in the Credit Product variable, we won’t be imputing those values instead we will be creating a new category as ‘Unknown’ for the variable.

	Age	Vintage	Avg_Account_Balance	Is_Lead	Credit_Product_na
Age	1.000000	0.631242	0.145232	0.230814	0.192168
Vintage	0.631242	1.000000	0.167433	0.279642	0.224488
Avg_Account_Balance	0.145232	0.167433	1.000000	0.053370	0.042413
Is_Lead	0.230814	0.279642	0.053370	1.000000	**0.531755**
Credit_Product_na	0.192168	0.224488	0.042413	0.531755	1.000000

Is Lead/Credit Product NA	0	1
0	183087	4350
1	33313	**24975**

From the above table, we can say that whenever the Credit Product is unknown, the proportion of the customer lead (Is Lead) is higher compared to other 2 categories (Yes and No).

3. Similarly, we also use the duplicated function in python to find any duplicate rows are present in the dataset and based on that, we must remove the duplicates using the drop duplicates function in python. We find that there are no duplicate rows present in the given dataset. So, we can now go ahead with the univariate and bivariate analysis.

Description of the Dataset

We will be seeing a basic description about the variables present in the dataset where we will be getting the no of unique values, count, maximum frequency, etc. for the categorical variables while we will be getting the basic measures like mean, minimum, maximum values, etc. for the numerical variables present in the dataset. We can refer to the below tables 4 and 5 for the description of the dataset for both categorical and numerical variables respectively.

Numerical Variables

	Age	Vintage	Avg_Account_Balance
count	245725.00	245725.00	245725.00
mean	43.86	46.96	1128403.00
std	14.83	32.35	852936.40
min	23.00	7.00	20790.00
0.25	30.00	20.00	604310.00
0.50	43.00	32.00	894601.00
0.75	54.00	73.00	1366666.00
max	85.00	135.00	10352010.00

Observations -

- **Age** - We have the age of the customers ranging from 23 to 85 which is a very wide range, and average age is around 44 and 75% customers have age less than or equal 54.
- **Vintage** - We have here the vintage variable which indicates the length of the customer relation in months. We have here the values ranging from 7 to 135 months

which means there are customers who are less than 1 year vintage as well around 10 - 11 years vintage.

- **Avg Account Balance** - Shows the Avg account balance of the customers in the last 12 months where we are having almost 75% customers have avg balance less than 13Laks while there is a maximum value of 1 crore.

Categorical Variables

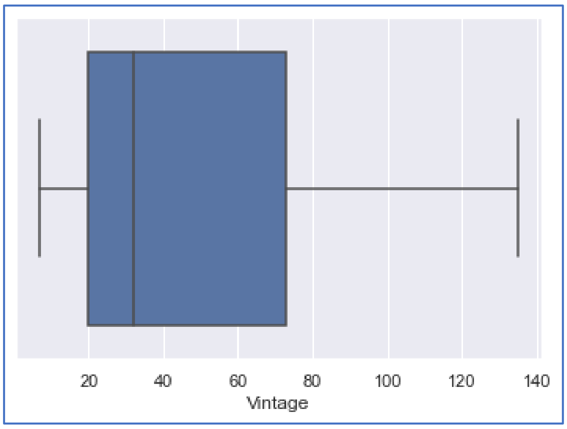
	count	unique	top	freq
Gender	245725	2	Male	134197
Region_Code	245725	35	RG268	35934
Occupation	245725	4	Self_Employed	100886
Channel_Code	245725	4	X1	103718
Credit_Product	245725	3	No	144357
Is_Active	245725	2	No	150290

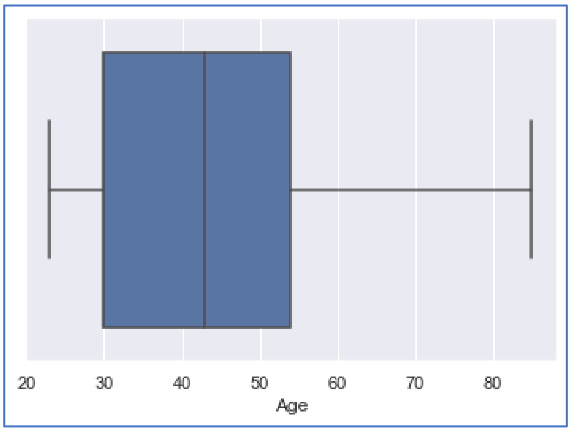
Observations -

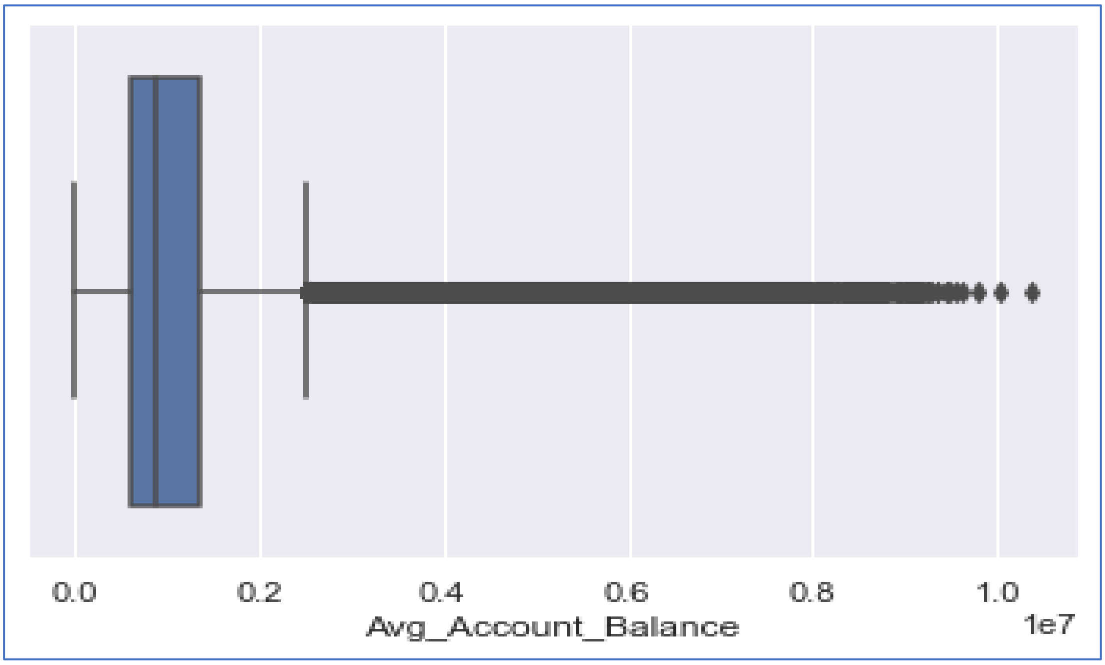
- **Gender -** Most of the customers are Males
- **Region Code -** Most of the records are from region code RG268.
- **Occupation -** Most of the customers are self-employed.
- **Channel Code -** Most of the customers channel code is X1.
- **Credit Product -** Most of the customers do not have an active credit product (Home loan, Personal loan, Credit Card etc.)
- **Is Active -** Most of the customers are not active in last 3 months.

Outliers Check

We can use boxplots to find whether there are outliers present in the dataset. It is shown in the figure 9 below. We find that there are no outliers present in our dataset except in the Avg Account balance variable where we have large number of outliers.







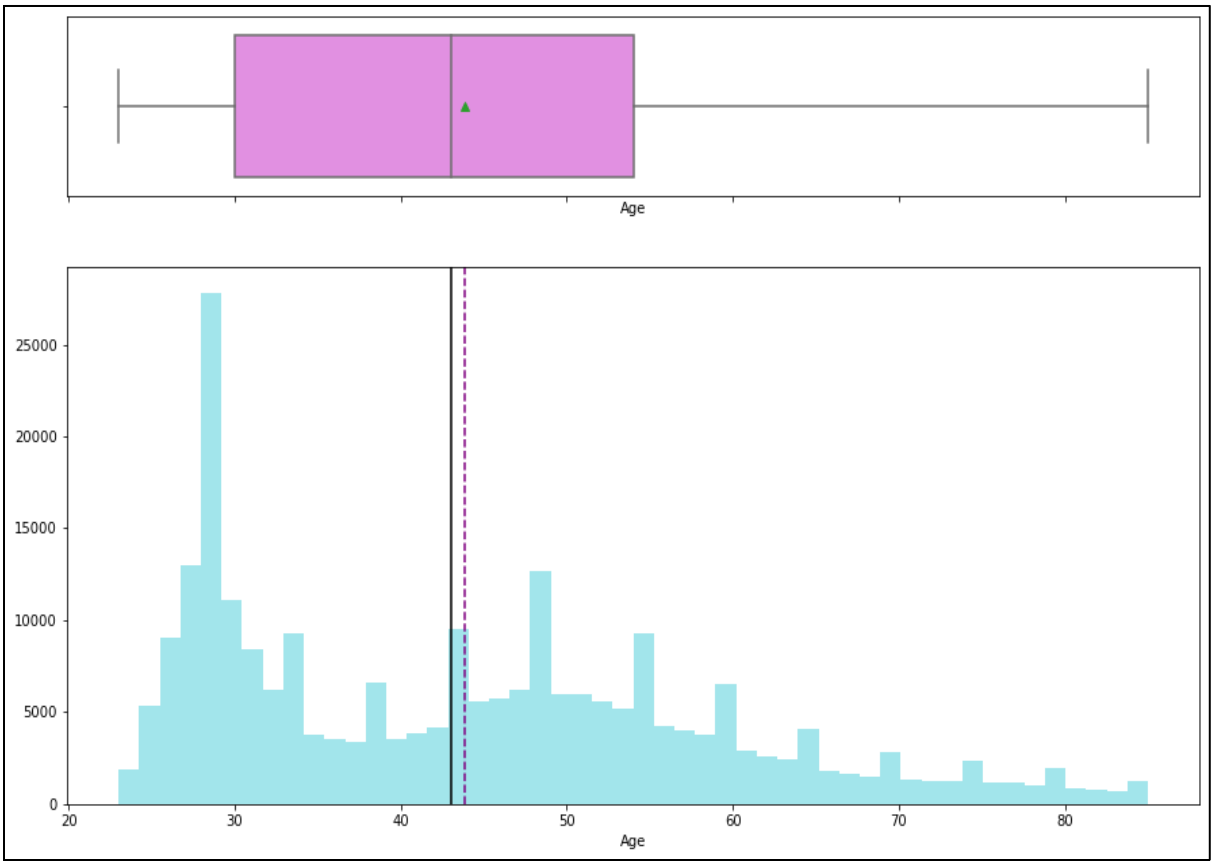
Here, we don’t have to drop or impute the outliers present in the Avg Account balance as we will lose important information and hence, we will proceed without replacing the outliers.

Univariate Analysis

Univariate analysis is used to analyse each individual variables present in the given dataset like the spread, distribution of the variable present in the dataset. We will be achieving this using the boxplots and histogram plots for the numerical variables. Similarly for the categorical variables, we will be using the count plots to determine the count of each category. We can see the plots from the figures shown below.

Boxplots and Histogram plots for Numerical Variables–

Age –



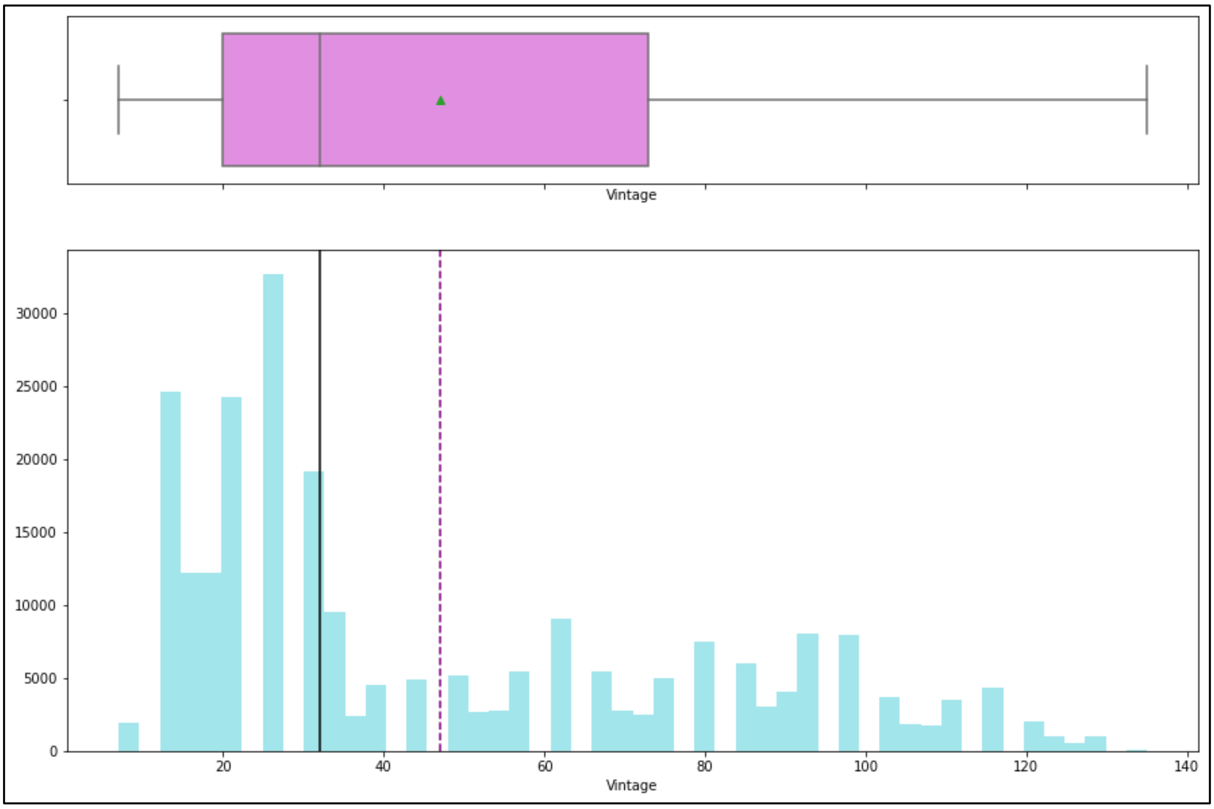
Observations -

- Age variable is right skewed.
- Mean and Median are close to each other.
- There are no outliers present in the age

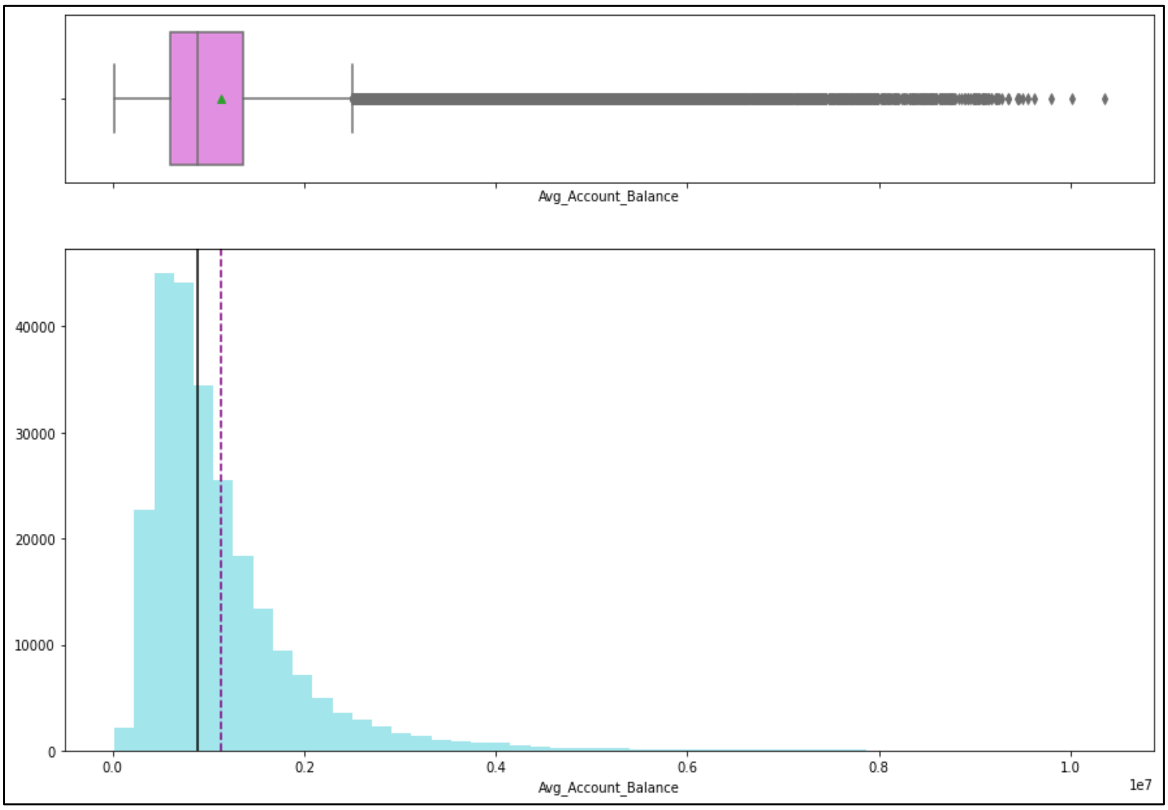
Vintage –

Observations

- Vintage variable is also right skewed.
- We get Median of around 30 months while mean is around 45-50 months, so there are some extreme values increasing the mean.
- Also, there are no outliers here.



Avg Account Balance -

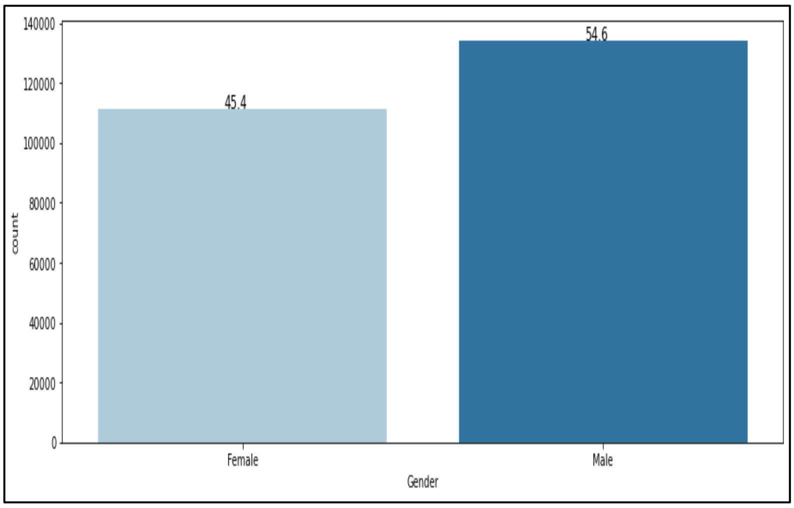


Avg Account Balance

- There are outliers present in Avg Account Balance.
- This variable is highly right skewed with many outliers.
- From the boxplot, we can that customers with more than 25 - 30 Lacs are outliers.

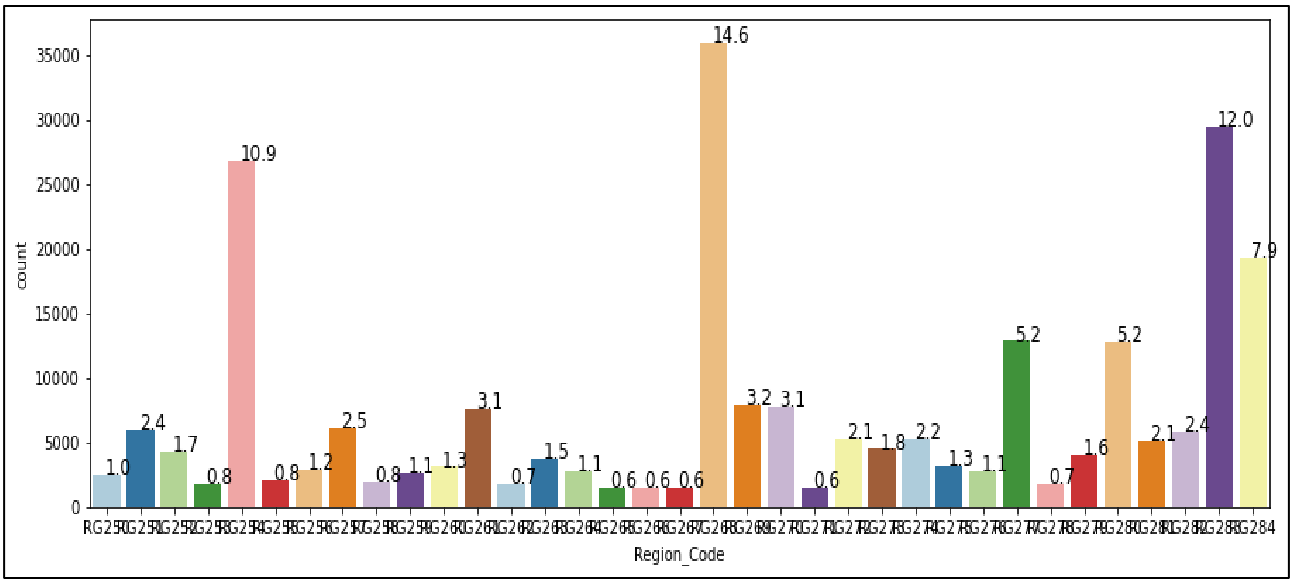
Count Plot for Categorical Variables

Gender -



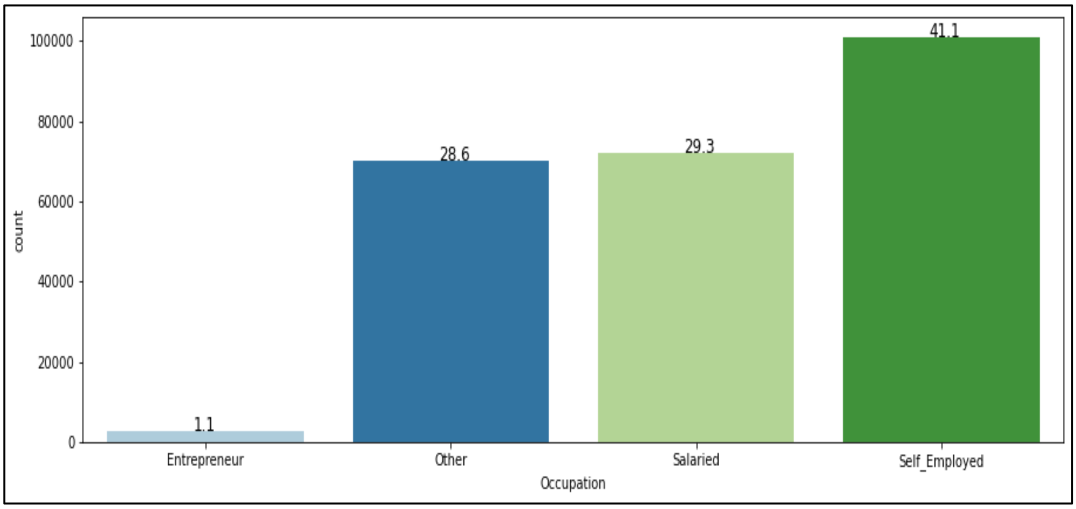
- We find that Male category customers are around 55%, and females are around 45%.

Region Code –



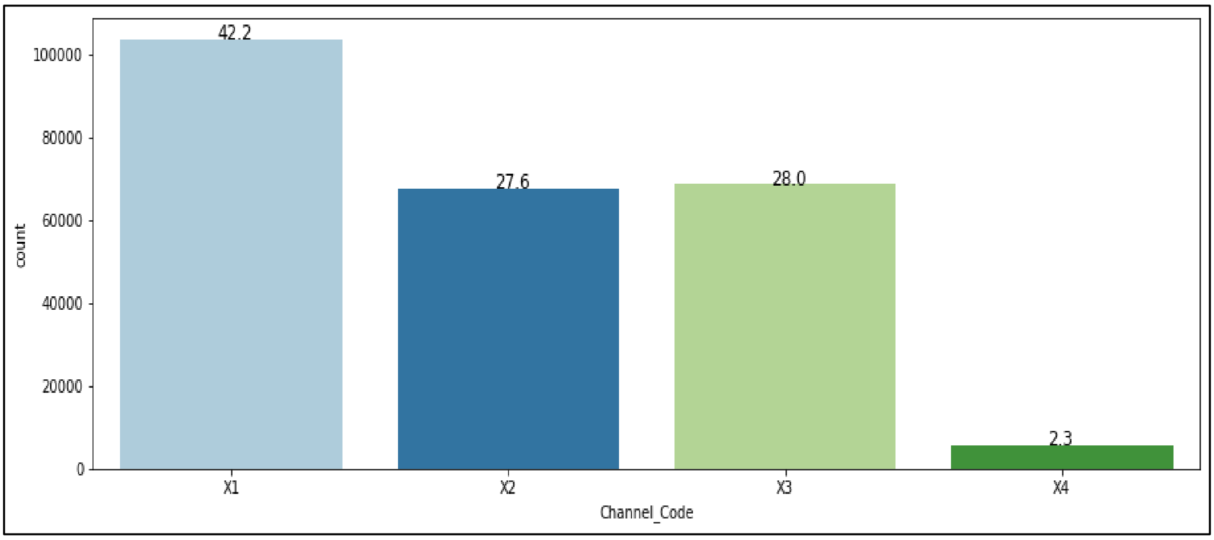
- We can see that up to 4 regions alone contribute to 45% of the customers.
- We can see many regions with less than 1% of the total customers.

Occupation -



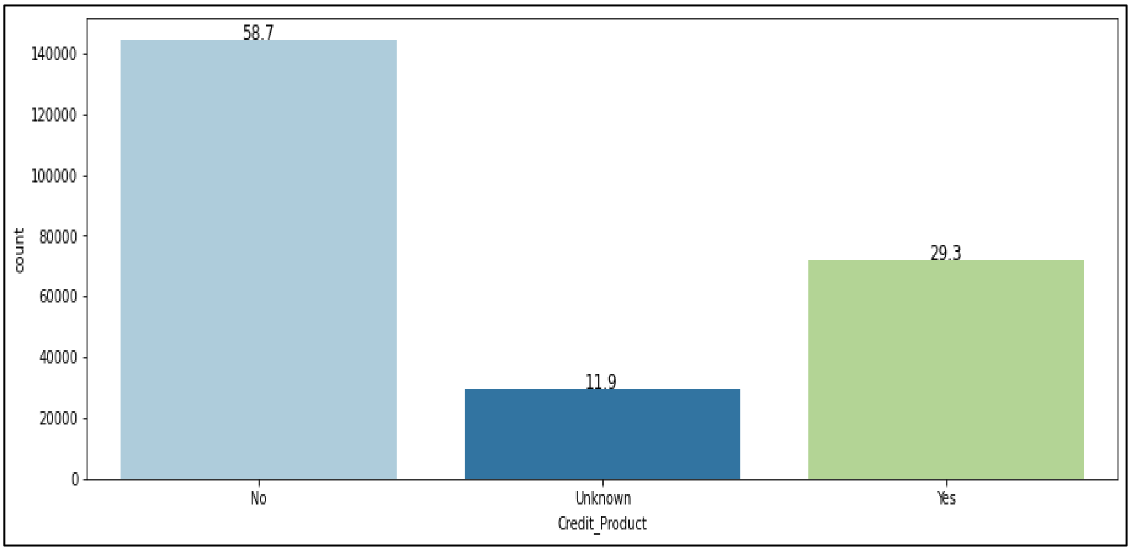
- Around 41% of the customers are self-employed while 30% customers are salaried and 1% belongs to Entrepreneur.

Channel Code -



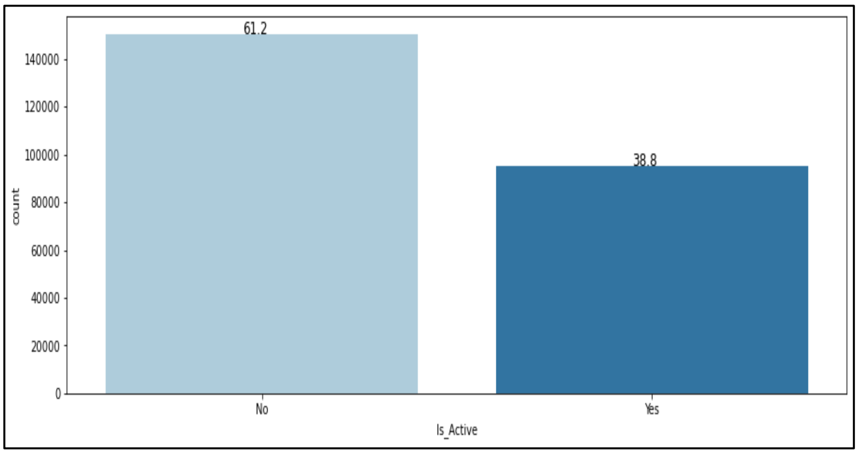
- 41% of the customers are from X1 channel code type while 28% are from X2 and X3 while 2% are from X4.

Credit Product -



- 59% of the customers don't have an active Credit Product while 29% have an active Credit Product.
- Here, Unknown means that we don’t know the customers status and it is of around 12%.

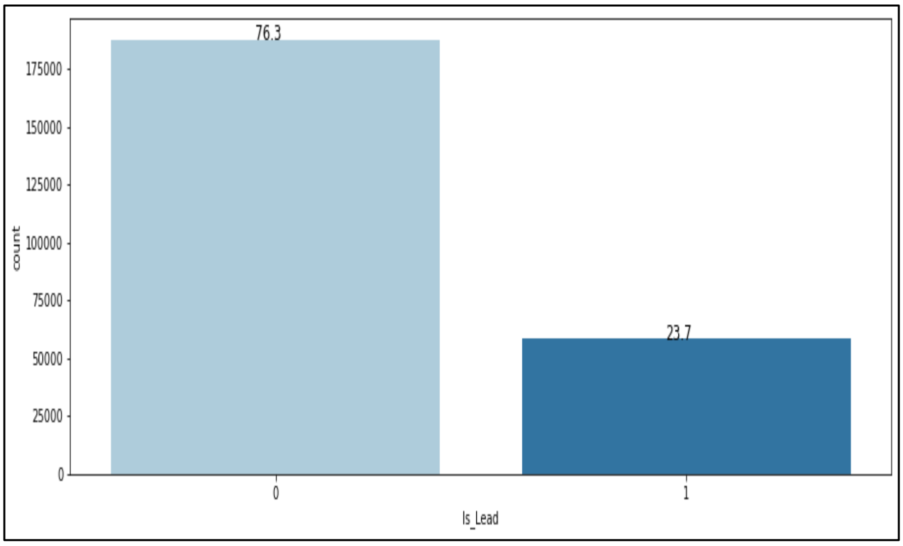
Is Active –



- 61% of the customers have not been active in the past 3 months while 39% have active.

Is Lead –

- This is the target variable. There is no imbalance in the given training set.
- We have 76% customers not interested in the credit card and 24% interested in the credit card.



Bivariate Analysis

Bivariate Analysis is used to determine the relationship between multiple variables present in the dataset. It is used to find useful insights about the variable's dependency on each other. This can be achieved with the help of correlation coefficients, heatmaps and pair plots.

Correlation Coefficient

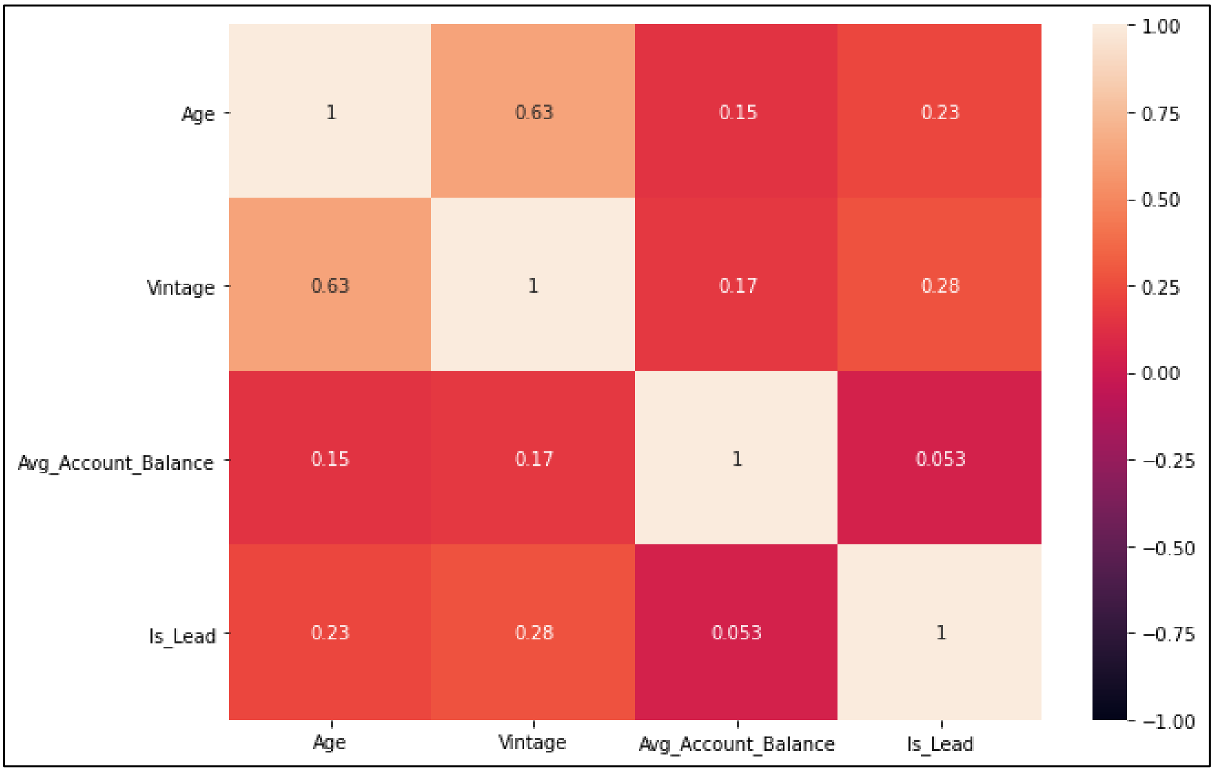
Correlation coefficient tells us about the strength of the relation between the numerical variables present in the dataset. Here, we will be using the corr function in python to generate the correlation coefficient from which we will be finding the strength of the relationship between the variables. We can see that here in the below table.

	Age	Vintage	Avg_Account_Balance
Age	1.000000	0.631242	0.145232
Vintage	0.631242	1.000000	0.167433
Avg_Account_Balance	0.145232	0.167433	1.000000

From the above table, we can see that there is not much higher correlation between the variables in the dataset. This can also be seen with the help of visualisations known as Heatmaps which will help us to find any correlation based on the correlation coefficient. It is shown in the below figure.

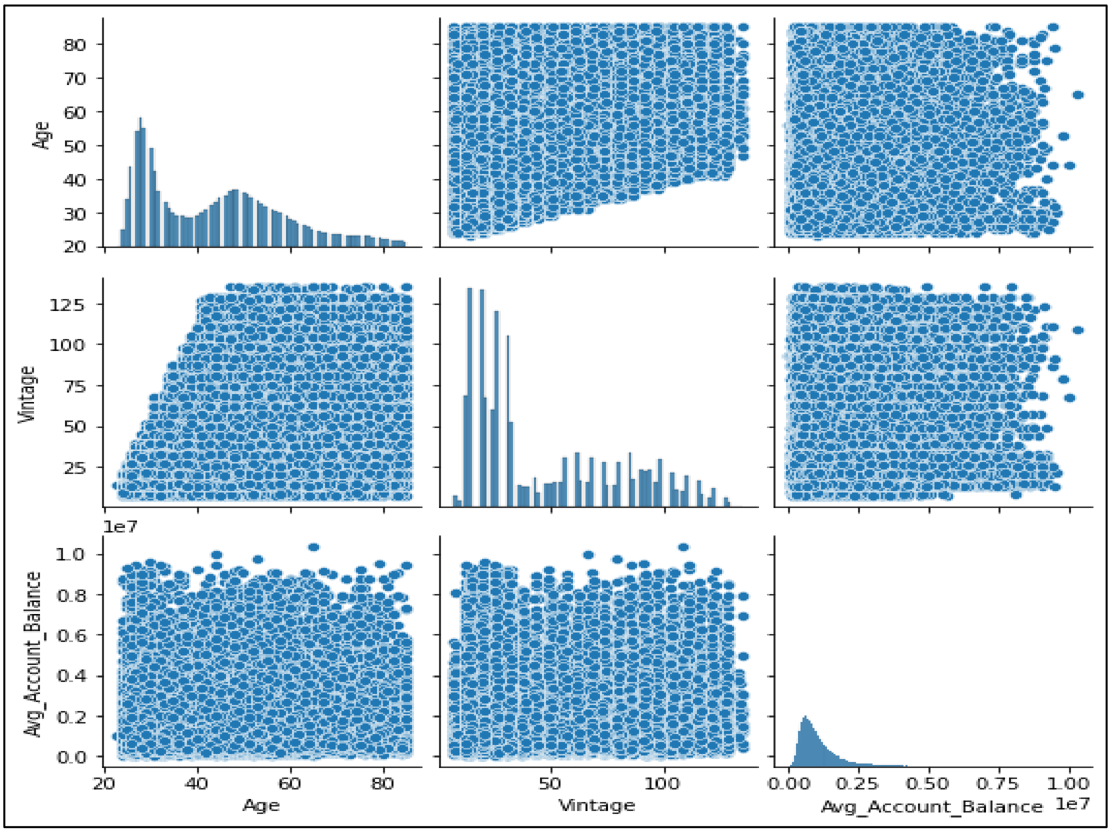
Heatmaps

From the heatmaps, we can clearly see that there are not much correlation present in the dataset between the variables in the dataset. So, we can go ahead with the further analysis of our bivariate analysis.



Pair Plots

Pair plots are used to determine the behaviour of the variable's relationship between each other with the help of scatter plots as shown below.



- There is no strong correlation either between the variables or with the target variable.
- There is small correlation for Age and vintage with Is lead variable.
- Also from the above pair plot, we can see that there is no relation between the numerical variables in the dataset.

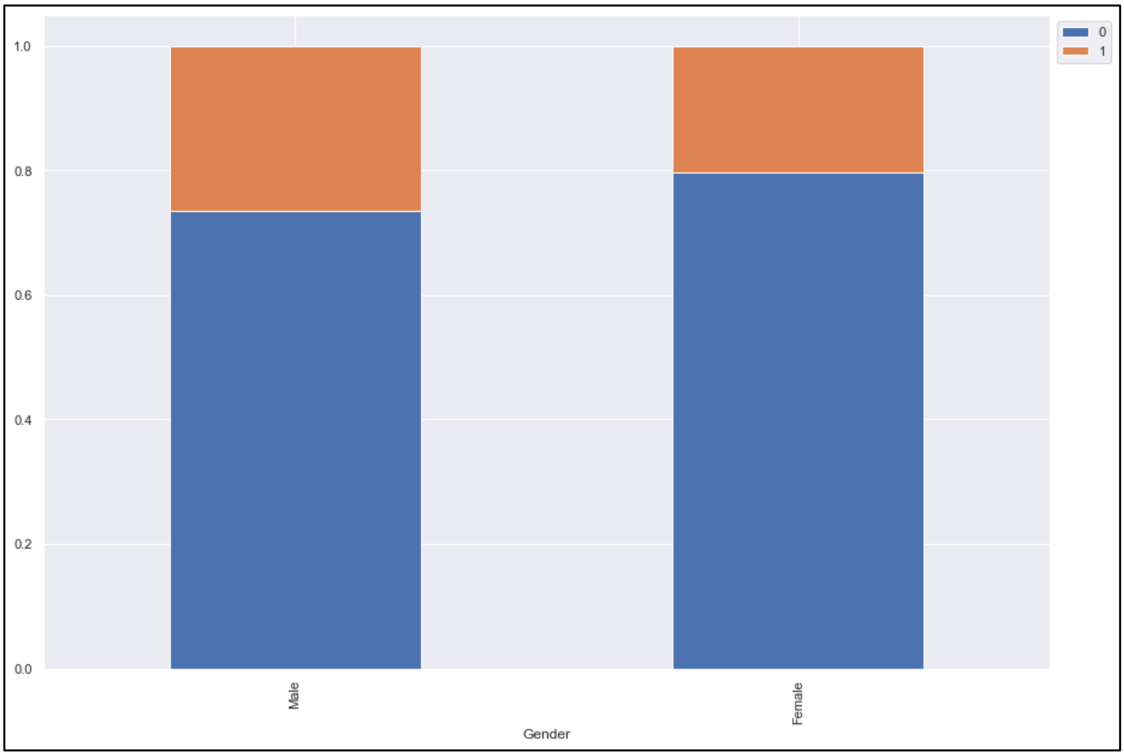
Bivariate Relationship between variables –

Now let us determine the relation between the Target variable and other variables using the boxplots, count plots and cross tab to find any insights if possible.

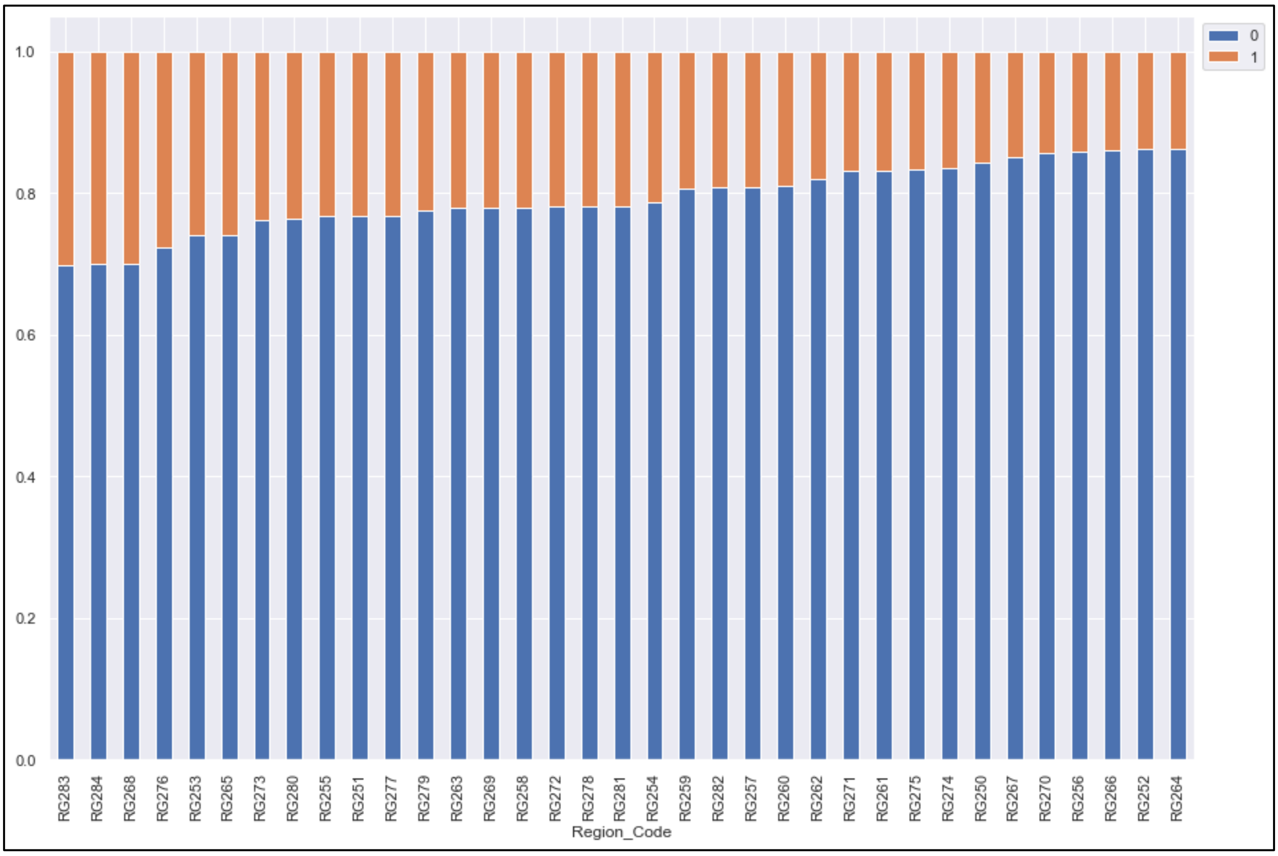
Categorical Variables Vs Target Variable

Gender Vs Is Lead

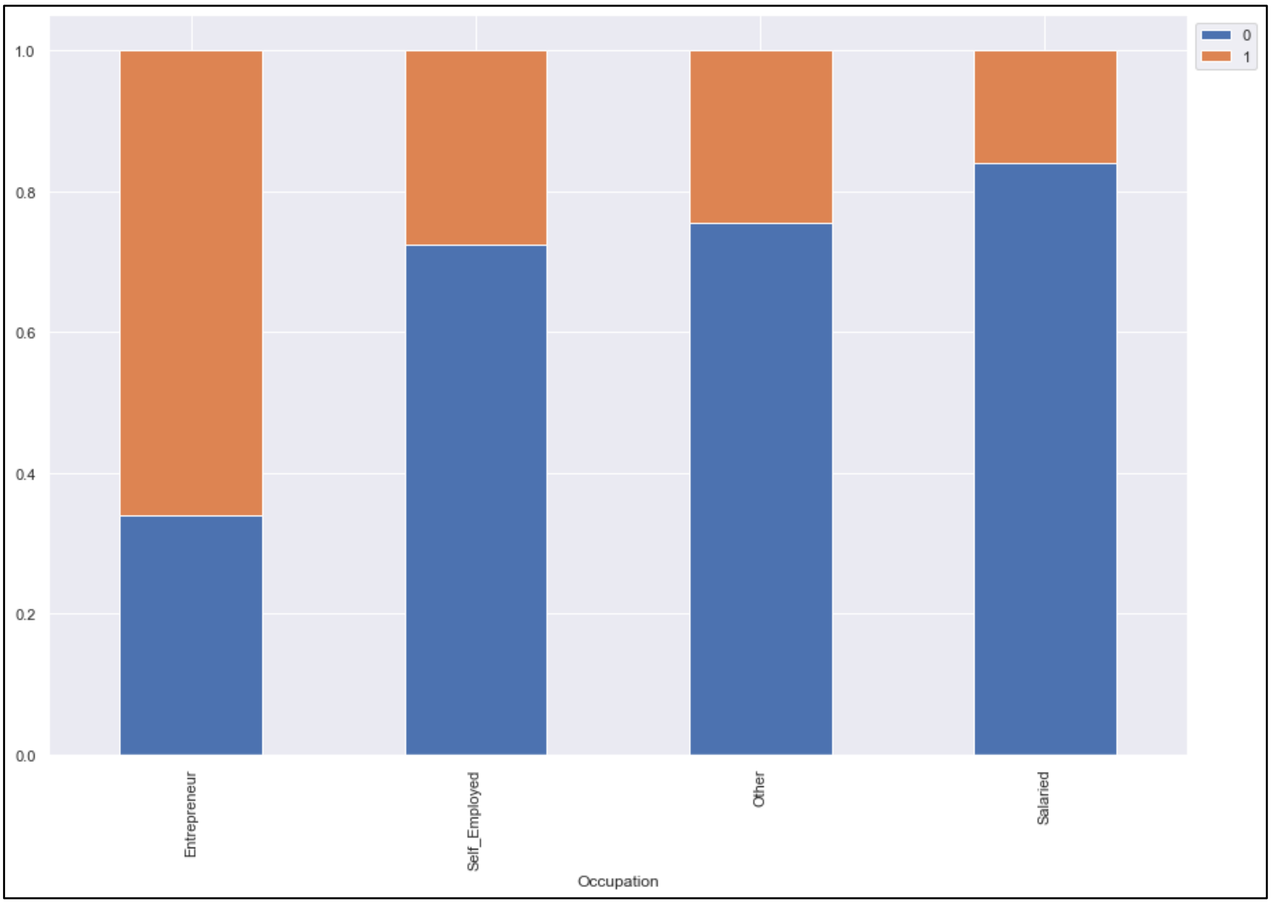
- From the below plot, Males are mostly likely to be interested in Credit card.
- While Females have less interest in the credit card compared to male.



Region Code Vs Is Lead



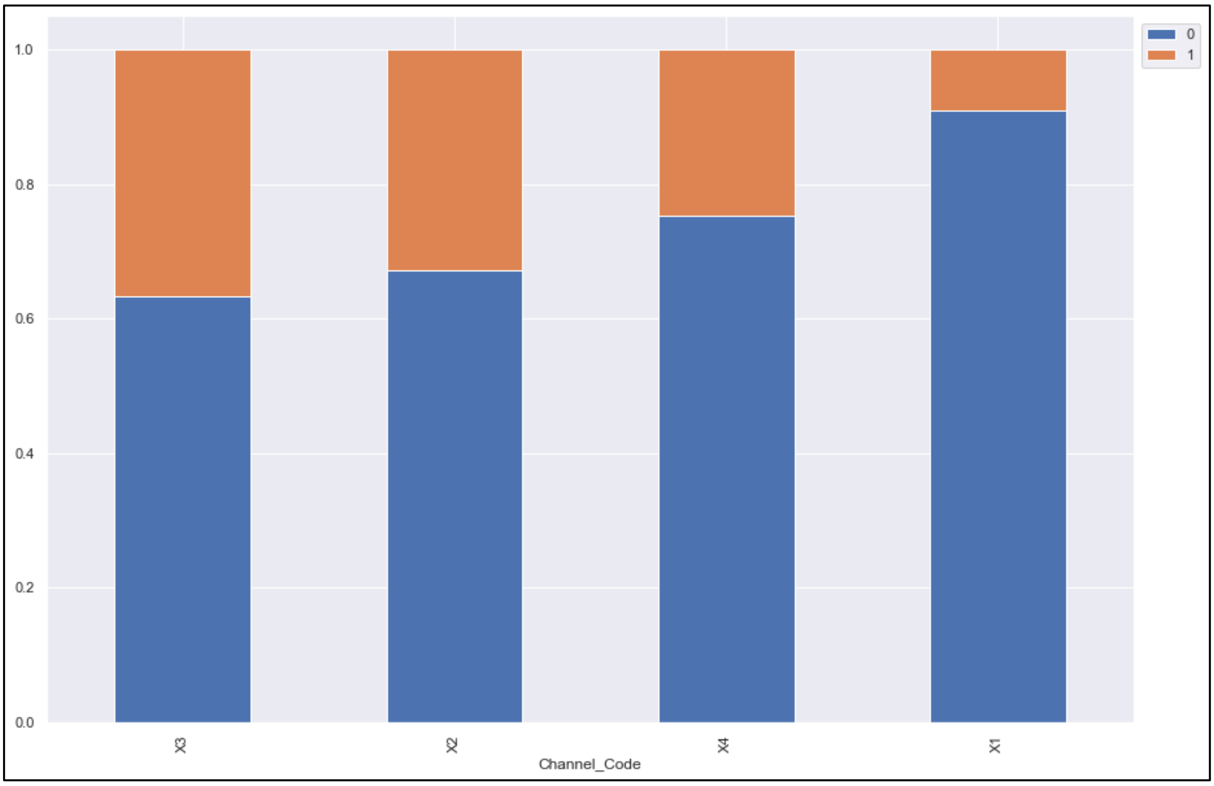
Occupation Vs Lead



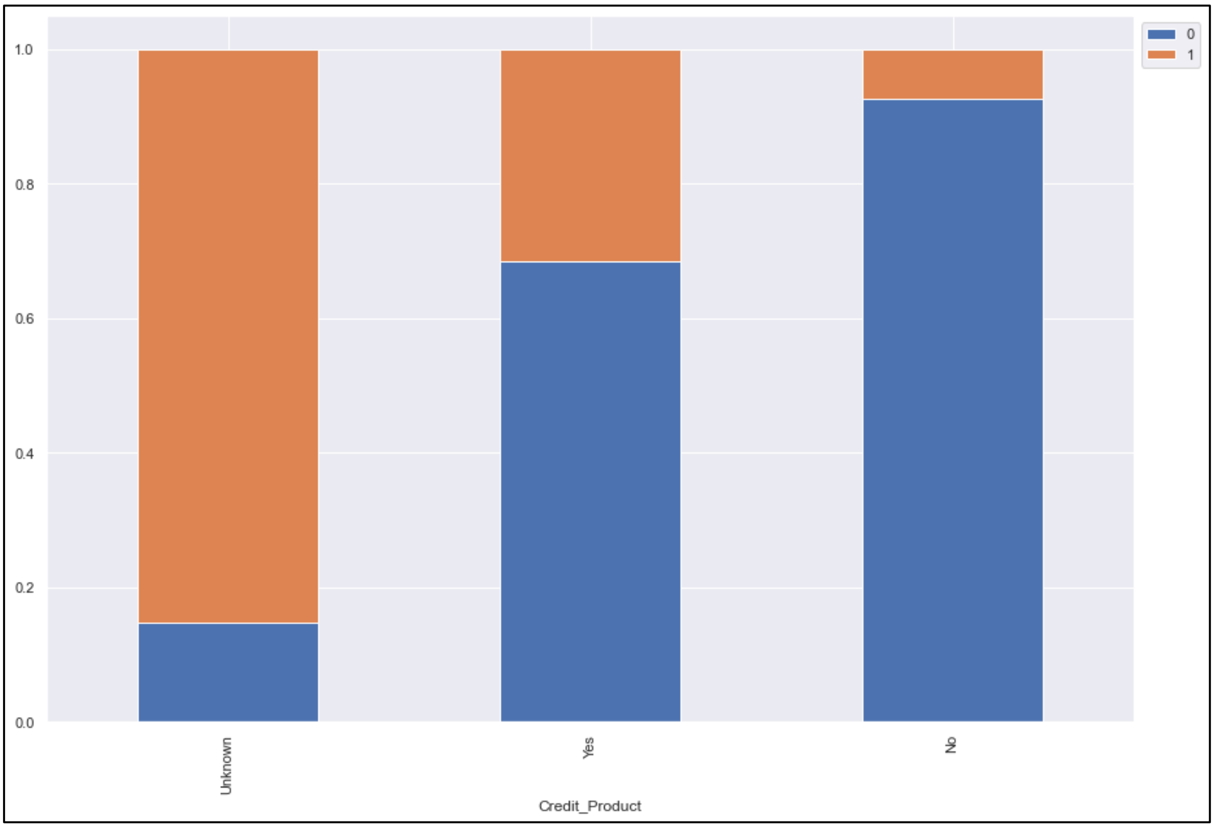
- Based on Occupation, Customers who are Entrepreneur are most interested in the credit card while customers whose occupation is salaried is least interested in credit cards.

Channel Code Vs Is Lead

- Even though most of the customers are from channel code X1, almost 90% of the customers from channel code X1 are not interested in Credit card.
- While customers from channel code X3,X4 and X2, customers from X3 channel code are most interested.

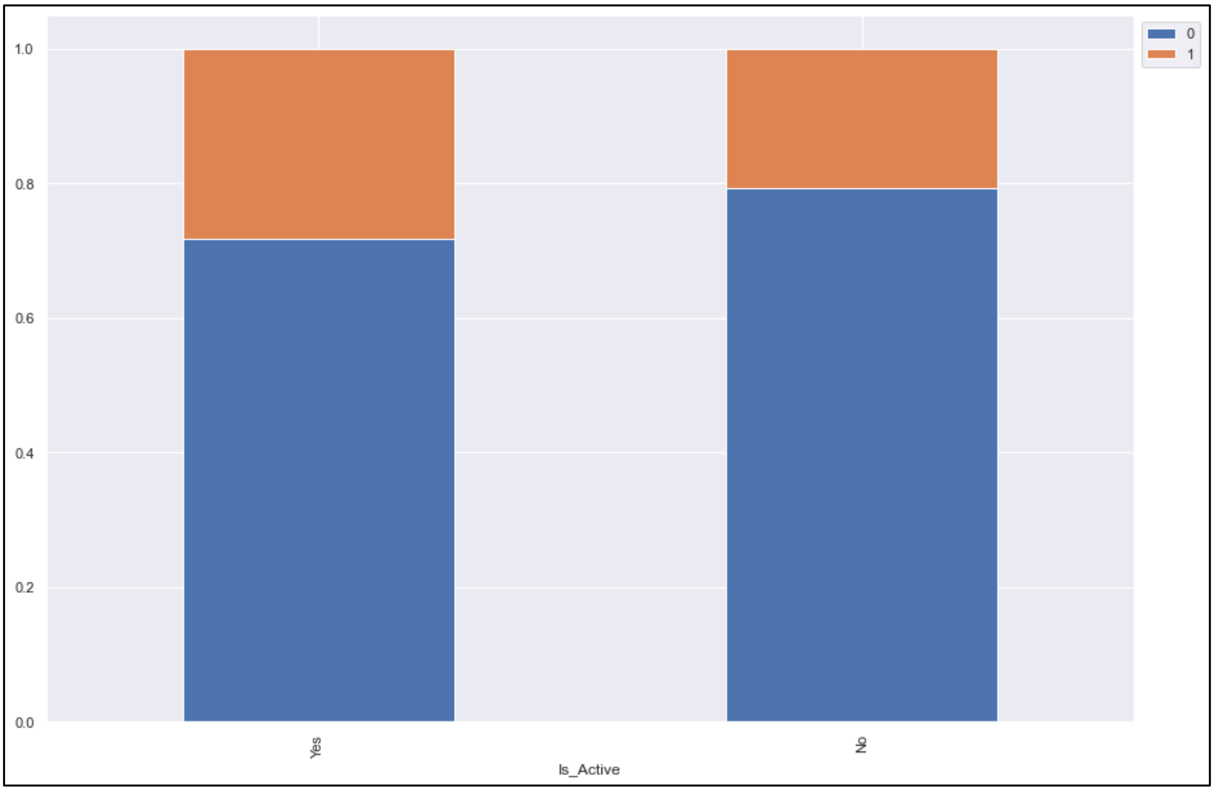


Credit Product Vs Is Lead



- Whenever the Credit Product is unknown, most of the customers are interested in the credit card.
- Similarly, when Credit Product is yes, customers are interested in credit card while when credit product is no, customers are not interested in credit card.

Is Active Vs Is Lead

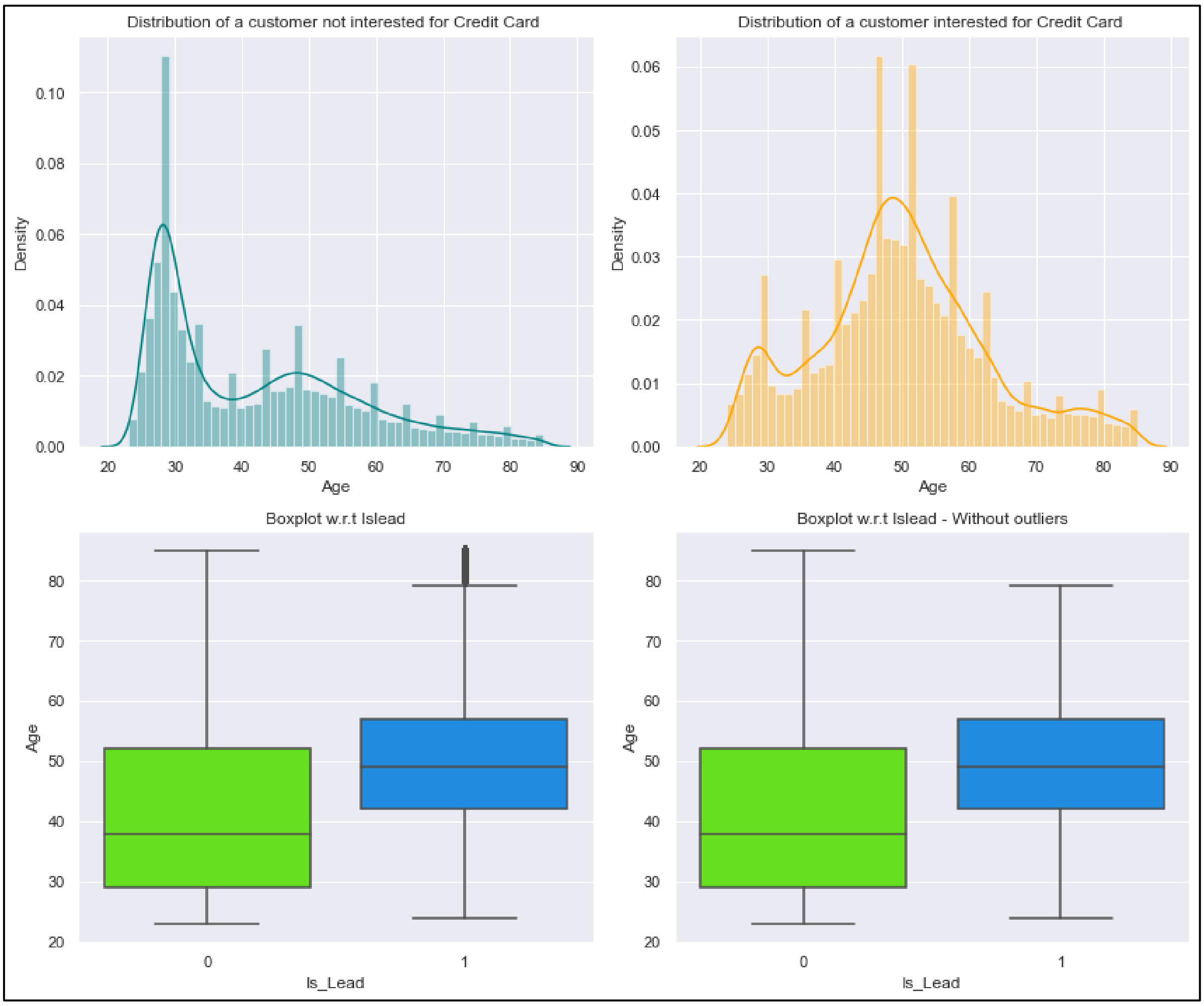


- From the Is Active variable, When the customers are not active, they are not interested in Credit Card.

Numerical Variables Vs Target Variable

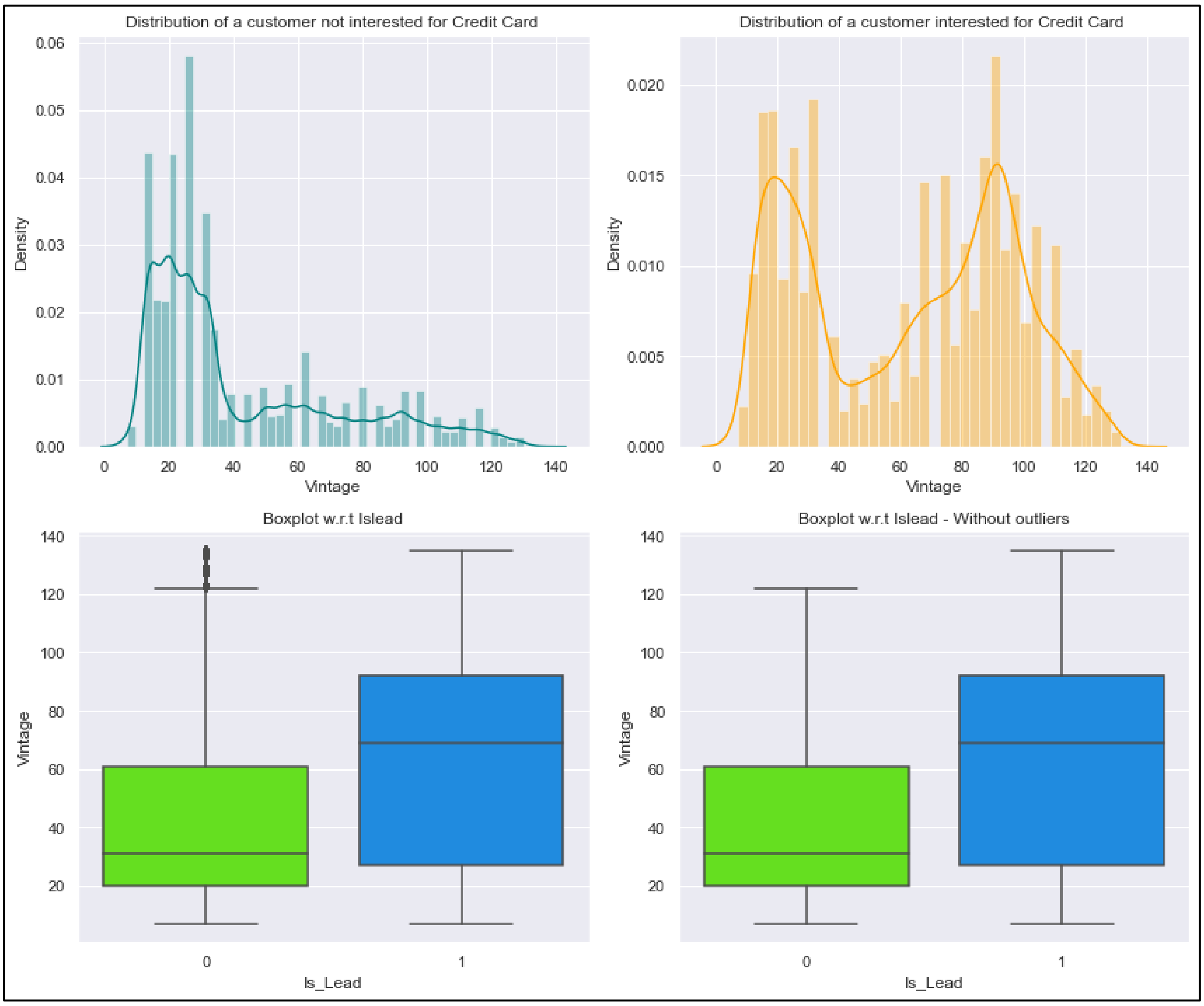
Now let us determine the relation between the Target variable and the numerical variables using the boxplot to find any insights if possible.

Age Vs Is Lead



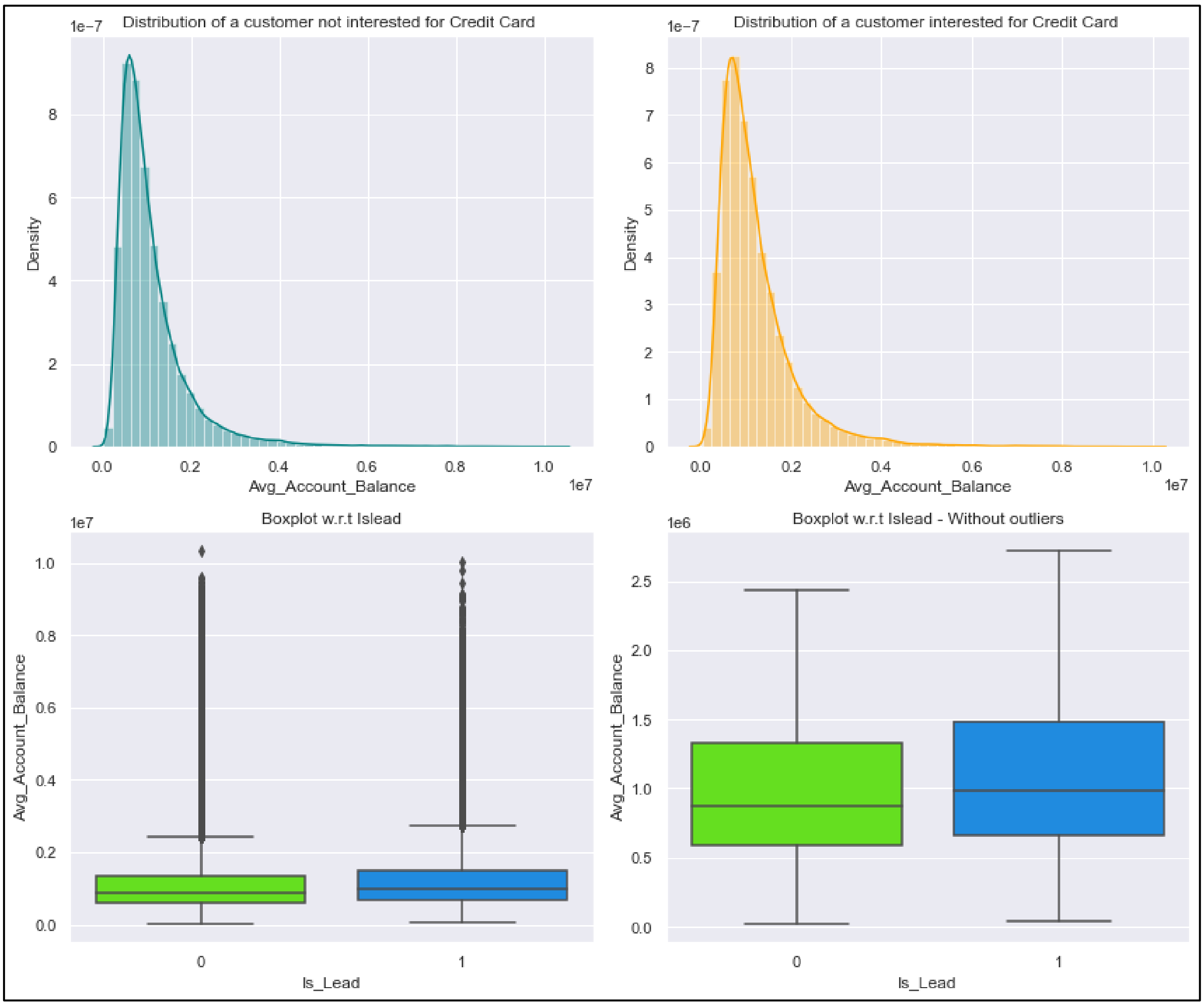
- Based on the Age, most of the customers interested in credit card belong to the age group 40 – 60.
- We can also see the mean of the customers interested in credit card at around 50 from the boxplots.
- Similarly based on age, most of the customers not interested in credit card belong to the age group 20- 40 and also the mean of the customers not interested is around 30-35.

Vintage Vs Is Lead



- Based on Vintage, the customers with more vintage are more interested in credit cards, also we cannot miss that some customers having vintage around 20-35 months are also interested in credit card.
- We can see from the mean of the vintage for the customers interested in credit card is near 70-80 months and higher vintage customers are most interested in credit card.
- Similarly, we can see the mean of the vintage for the customers not interested in credit card is near 30 months and lower vintage customers are not interested in credit card.

Avg Account Balance Vs Is Lead



- Not much can be said based on the Average account balance as both the set of customers interested as well not interested in credit card are having nearly same mean.

Model Building Pre-Processing –

The proposed model in the project is based on the prediction models as the dependent variable is nominal and takes the values of 0's and 1's indicating whether the customer is showing higher intent towards a recommended credit card, where '0' represents customer does not show interest and '1' represents the customer shows interest for the credit card. We have applied Logistic Regression, Decision Trees, Random Forest, Naïve Bayes, Boosting and

Bagging and the final model was developed with one of the models done below. The model validation was performed using Accuracy, AUC, precision, recall and f1 scores.

Model Building

Method 1 -

Once we are done with extracting the Target variable and independent variables separately, we will be splitting the data set into train and validation set for us to build the model on the train set and test the model on the validation set.

All the pre-processing steps like scaling, encoding, Missing values imputation, outliers' imputation is performed parallelly on both the train set and test set given in the problem.

Encoding Process –

We will be using One Hot encoding process to encode the categorical variables as the ML models will not be able to understand the categories or categorical variables. This can be done using get dummies function in python to create dummy variables for each category in the variable.

Normalisation and Scaling –

We will be using the log transformer and the Standard scaler function to scale the numerical variables to get into the same scale so that the ML model performs better and gives us accurate results.

Model Imbalance -

Is Lead (Target Variable)	Proportion
0	76.28

1	23.72

We don’t have any imbalance in our target variable, so we can proceed with the model building process.

Models Built –

Model Name	ROC AUC Score - Train	ROC AUC Score - Validation
Logistic Regression	0.859	0.857
KNN Model	0.899	0.841
LDA Model	0.858	0.856
Decision Tree Classifier	0.884	0.868
Random Forest	0.880	0.871
Gradient Boosting	0.904	0.870

Method 2 -

Once we are done with extracting the Target variable and independent variables separately, we will be splitting the data set into train and validation set for us to build the model on the train set and test the model on the validation set.

All the pre-processing steps like scaling, encoding, Missing values imputation, outliers’ imputation is performed parallelly on both the train set and test set given in the problem.

Encoding Process –

We will be using One Hot encoding process to encode the categorical variable Region code and label encoding for all the other categorical variables as the ML models will not be able to understand the categories or categorical variables. This can be done using get dummies function in python to create dummy variables for each category in the variable.

Normalisation and Scaling –

We will be using the Min Max scaler function to scale the numerical variables to get into the same scale so that the ML model performs better and gives us accurate results.

Models Built –

Model Name	ROC AUC Score - Train	ROC AUC Score - Validation
Logistic Regression	0.729	0.726
LDA Model	0.726	0.723
Decision Tree Classifier	0.893	0.861
Random Forest	0.879	0.869
Gradient Boosting	0.904	0.870

Conclusion –

Final model is selected based on the AUC score shown above and we can see that in both the approaches used, the Gradient Boosting model performed better than other models and results in classifying both the classes better.