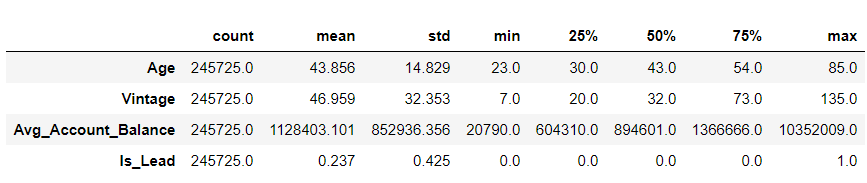
**Approach Used:**

1. **Understanding the data:**

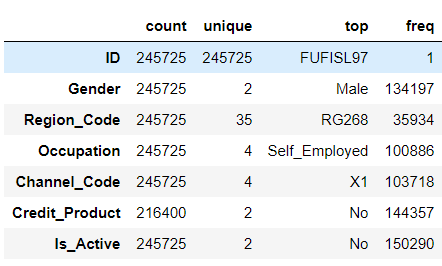
We have 4 columns with data type int and 7 with data type object/string. Only one column (credit\_product) has missing values.



Mean age is 43.856 and median age is 43, i.e. approximately equal. And maximum age value is 85 years and minimum is 23 years. Standard deviation for age is 14.829.

Vinage has mean 46.959 months while median 32 months. Maximum vintage value is 135 months and minimum vintage value is 7 months. Standard deviation for vintage is 32.353.

Average account balance has mean 1128403.101 and median 1366666.0. Standard deviation for Avg. Account balance is 852936.356.



We have higher numbers of males (Male: 134197)

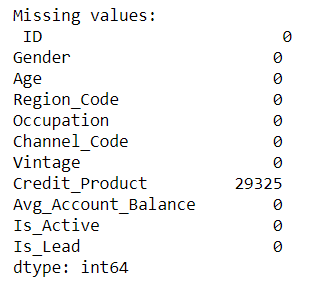
High number of people are not active, ie. 150290.

There are 35 regions and RG268 has top frequency = 35934.

Occupation categories are 4 and Self\_Employed has highest frequency : 100886.

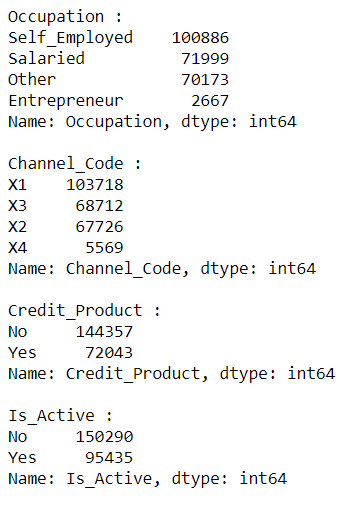
Channel Codes have 4 categories and X1 has highest frequency, i.e. 103718.

Credit Product are of 2 type and most of people has no active credit product. Frequency is 144357.



There are 29325 missing values in Credit\_product.

There is no duplicate record in the dataset.



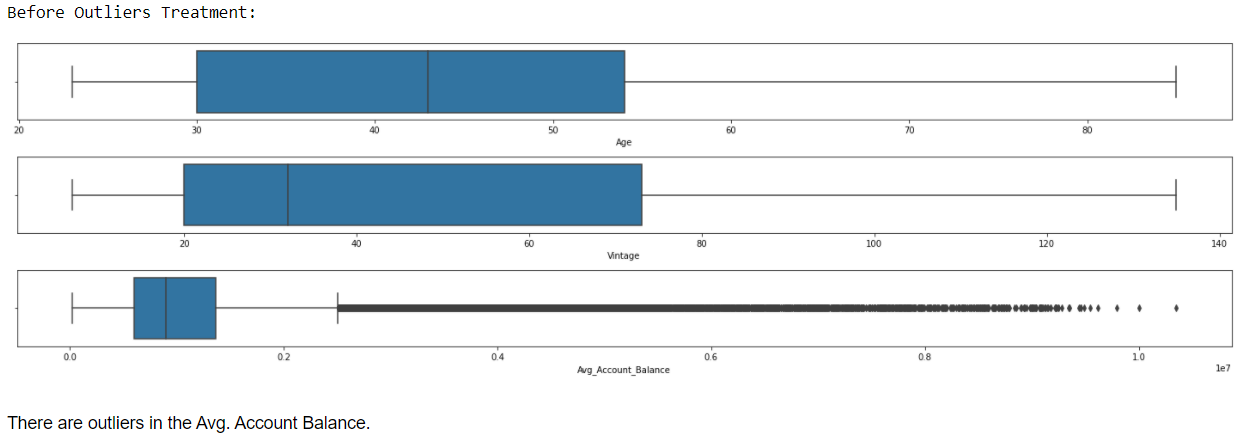
From the above analysis, it is clear that there is no missing or unwanted values in the category on the dataset.

Region\_code has most categories.

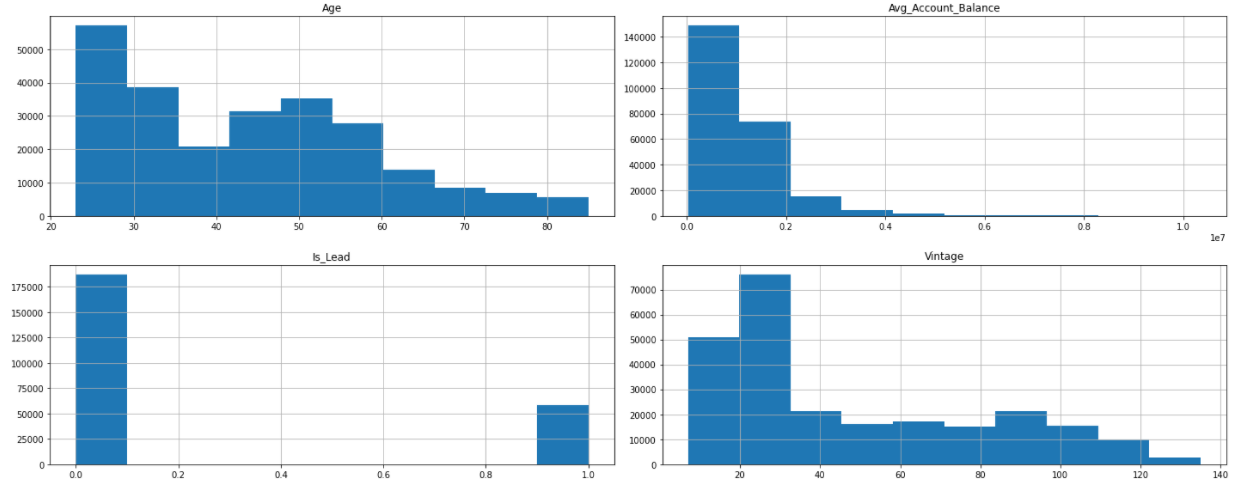
All the category columns have string values and we need to convert them to ordinal before model building

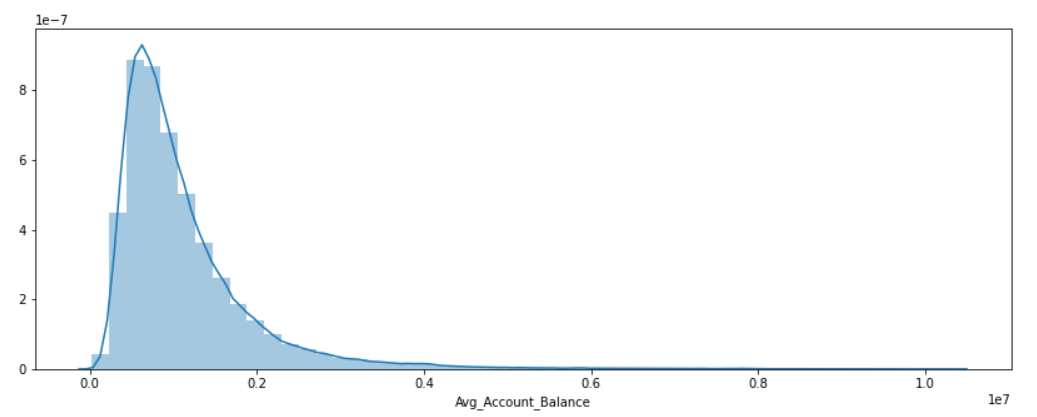
Univariate, bivariate and multi variate analysis

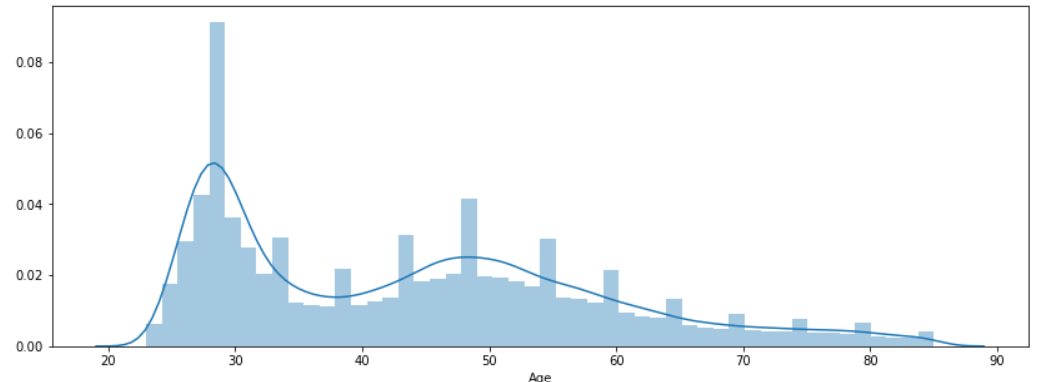
Before doing one hot/ label encoding, lets do univariate, bivariate and multi variate analysis to understand the data properly.

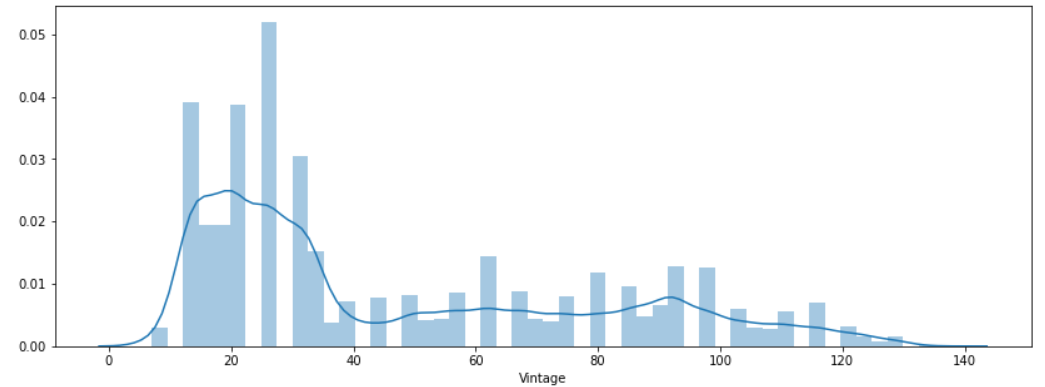


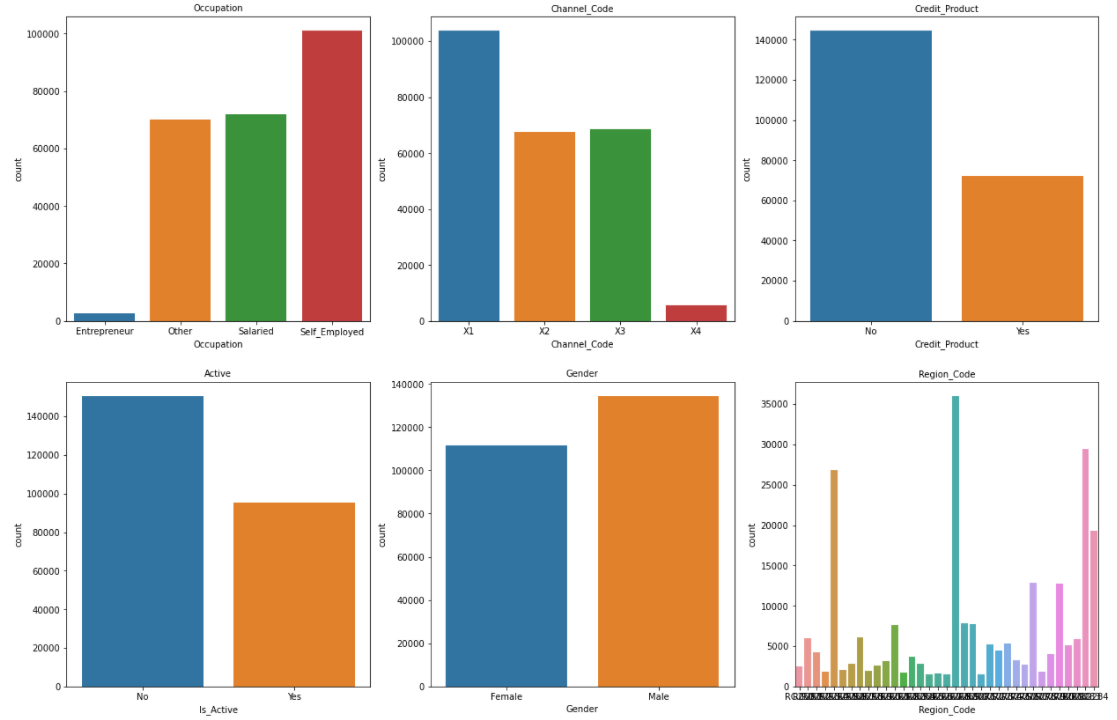
The data is not normally distributed.





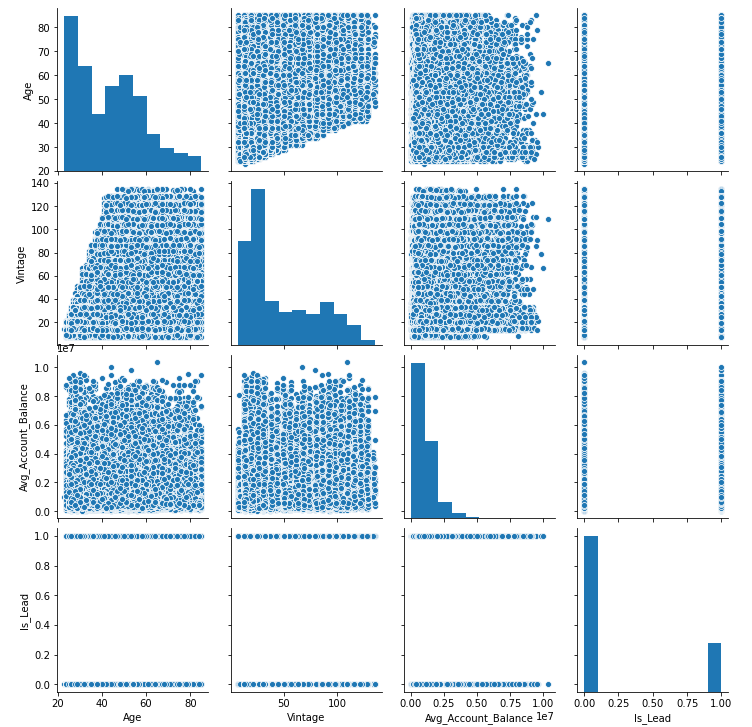






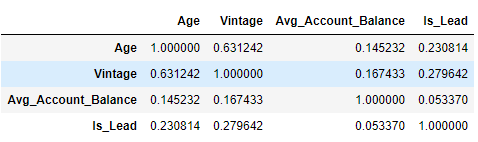
Most of the candidates are Self\_Employed.

Channel\_Code X1 has most records. There are fewer people who have opted for Credit\_product and who are Active.

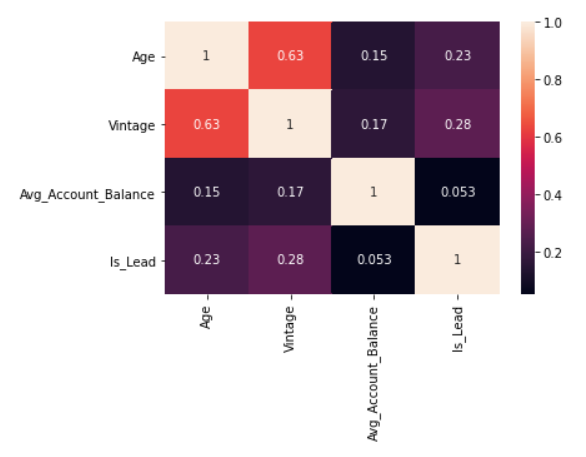


There is no clear relationship between the variables.

Correlation Matrix:

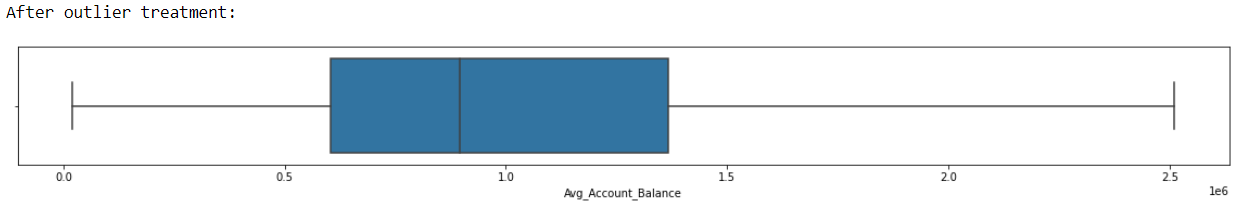


Correlation heat map:

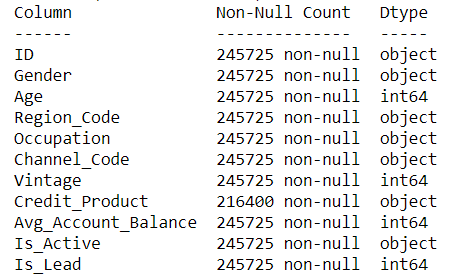


1. **Outlier treatment:**

There are outliers in the Avg. Account Balance. I have used the IQR to treat the outliers.



1. **Missing data handling**

****

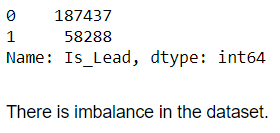
There are missing values in the Credit\_Product column. There are 29325 missing values in Credit\_product.

1. **Scaling the Data**

The data gave different scales like Age is measured in years, vintage is measured in months, and Avg\_Account\_Balance is money. So, we need to scale the data before model building. I have used zscore.

1. **Checking imbalance in the data**

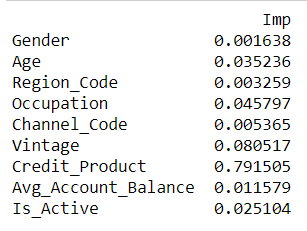
There is imbalance in the dataset and I have used oversampling to treat the imbalance before model building.

****

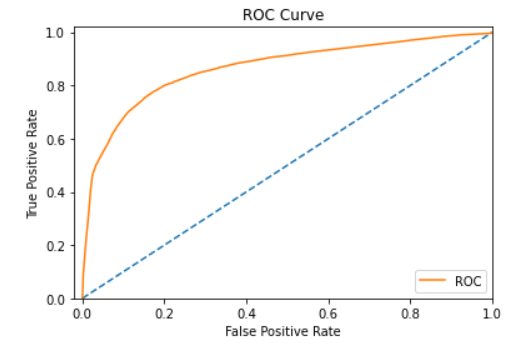
1. **Model building & Hyper tuning**

I have used Decision tree, Random Forest, Neural Networks, KNN, Logistic regression, Bagging Classifier, Ada Boost Classifier, and Gradient Boosting Classifier for building different models and used Grid Search CV for hyper-tuning the models.

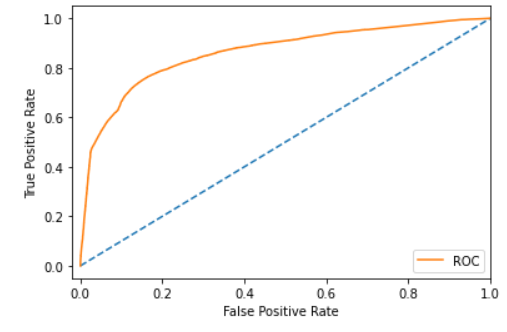
Feature Importance:



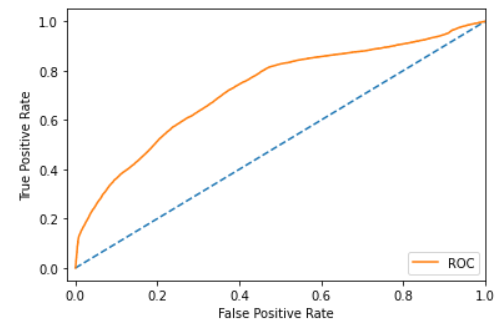
Decision Tree: AUC: 0.8650053210628199



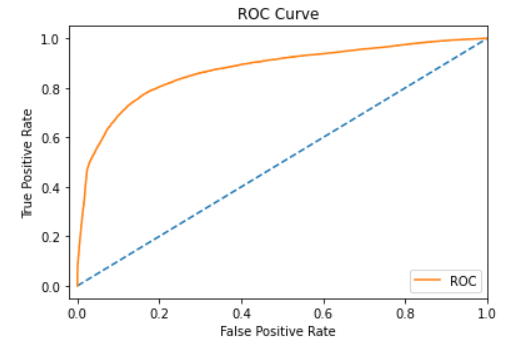
Random Forest Classifier: AUC: 0.8621361239812824



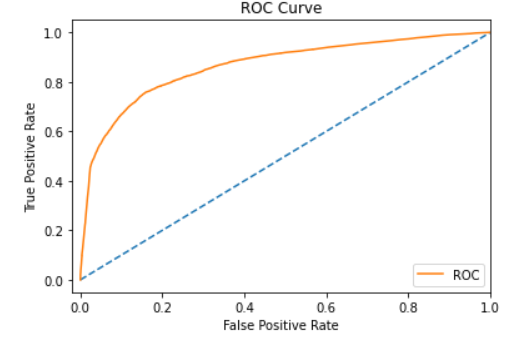
Logistics Regression classifier: AUC: 0.7247593889291688



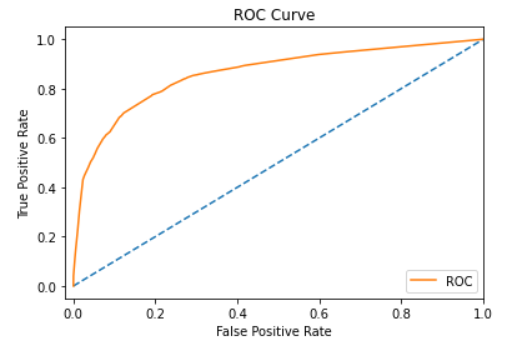
Bagging Classifier: AUC: 0.8714722264269885



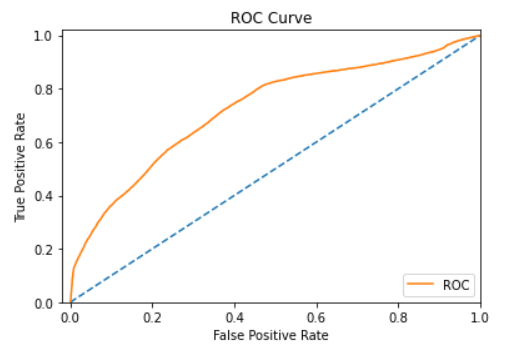
AdaBoostClassifier: AUC: 0.864738987774871



GradientBoostingClassifier: AUC: 0.8609702572705915



KNeighborsClassifier: AUC:0.8231



MLPClassifier: