**ASSIGNMENT REPORT:**

**OBJECTIVE:** Customer Churn Prediction for Telecom Company.

**BUSINESS PROBLEM:** The company is concerned about customer churn and wants to develop a classification model to predict which customers are likely to churn. As a data scientist you are required to analyse the data and implement a Machine Learning Algorithm for solving the classification problem.

**DATASET DESRIPTION:**

The dataset includes customer information such as demographics, usage patterns and scores associated with the profile of customer.

The dataset contains 14 columns of which 13 columns can be taken as Input features for our data analysis , one column is our target variable or output variable namely ‘Exited’.

In dataset we have two types of columns numerical and categorical.

There are particular columns which will not help us in analysing the problem statement they are , 'RowNumber','CustomerId','Surname'.

The numerical columns are:

col\_numerical = ['CreditScore','Age','Balance','EstimatedSalary','Exited']

The categorical columns are:

col\_categorical = ['Geography','Gender','Tenure','NumOfProducts','HasCrCard','IsActiveMember','Exited']

The Target columns ‘Exited’ contains two variables 1 or 0 , where 1 means The customer has left the firm, and 0 means the customer has not left the firm.

**CONSTRAINTS:**

On briefly analysing the dataset and problem statement , we can observe some constraints and preprocessing needs for our dataset.

1. The problem statement requires classification Machine Learning Algorithms.
2. Need to remove outliers from the numerical columns so as to get better accuracy and get appropriate regularization for model training.
3. The categorical columns are required to get One Hot Encoded.
4. Biased Data can result in biased predictions and unfair outcomes, so we are required to get train and test data splits with almost equal proportions of both outcomes.
5. Hardware or Computational Resource can be an issue in Training models with more hyperparameters, these can result in Ensemble and Boosting Algorithms.

**PERFORMANCE METRICS:**

* **Accuracy:** Measures the overall correctness of a model by calculating the ratio of correctly predicted instances to the total instances.
* **F1 Score:** Harmonic mean of precision and recall, providing a balanced metric for binary classification, particularly useful when classes are imbalanced.
* **Precision:** Proportion of true positive predictions among all positive predictions, indicating the model's ability to avoid false positives.
* **Recall:** Proportion of true positive predictions among all actual positives, revealing the model's ability to capture all relevant instances.
* **AUC-ROC Score:** Area Under the Receiver Operating Characteristic curve, quantifying the model's ability to distinguish between classes, with higher values indicating better performance.

**MACHINE LEARNING ALGORITHMS:**

The Machine learning Algorithms which can used in this problem statement are:

* **K-Nearest Neighbors (KNN):** A simple and intuitive algorithm where predictions are based on the majority class of the k-nearest data points in the feature space, making it effective for classification and regression tasks.
* **Logistic Regression:** A linear model used for binary classification, estimating the probability of an instance belonging to a particular class. It's particularly interpretable and works well when the relationship between features and the log-odds of the response is approximately linear.
* **Decision Tree:** A tree-like model where decisions are made by splitting data based on feature conditions, leading to a tree structure. It's easy to understand and interpret, suitable for both classification and regression tasks.
* **Random Forest:** An ensemble model consisting of multiple decision trees, each trained on a random subset of the data. It improves accuracy and generalization by reducing overfitting and capturing diverse patterns in the data.
* **AdaBoost:** An ensemble method that combines weak learners sequentially, giving more weight to misclassified instances in each iteration. It adapts and focuses on difficult-to-classify cases, improving overall model performance.
* **XGBoost (Extreme Gradient Boosting):** An efficient and scalable implementation of gradient boosting, XGBoost builds a strong ensemble of decision trees sequentially, optimizing for predictive accuracy. It incorporates regularization and parallel processing, making it powerful for various machine learning tasks.Top of Form

**PROCEDURE:**

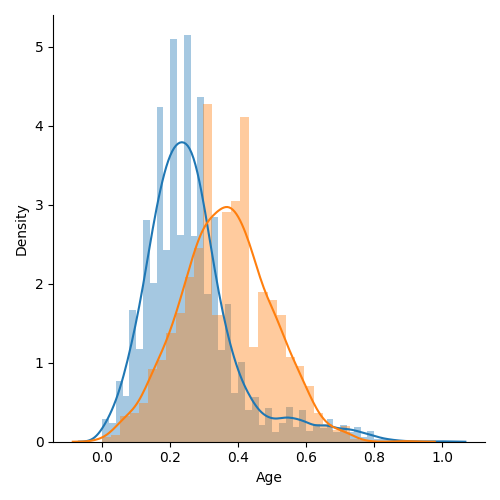
1. Importing required libraries on Jupyter Notebook.
2. Importing The dataset (“Churn\_Modelling.csv”) as pandas dataframe ‘df’.
3. Removing irrelevant or not important features such as 'RowNumber','CustomerId','Surname'
4. Split the dataset based on datatype category of numerical and categorical.
5. We are required to normalise the numerical columns , these is done using MinMaxScaler imported from sklearn.prerpocessing library.
6. Carry Out EDA(Exploratory Data Analysis) on the numerical data , we will plot
   1. Distribution Plot
   2. Box Plot
   3. Histogram
   4. Violin Plot
7. Observe The plots against the target variable and find out any insights from the data.
8. We are required to OneHot Encode the categorical columns, these is done using a pandas module get\_dummies() .
9. Carry out EDA for categorical columns, we will plot graphs against categorical columns and the target variable.
10. Bar Plots
11. Pie Charts
12. Observe the Plots against the target variable and find out any insights from the data.
13. Join the df\_numerical and df\_categorical as df\_preprocessed.

**Exploratory Data Analysis (EDA):**

After plotting the graphs with the help of Seaborn and Matplotlib Library in python, we can observer some insights in the data with respect to the target variable.

The features which showed some significant insights are :

AGE :



Observations:

We can observe there is a clear difference between the two groups,

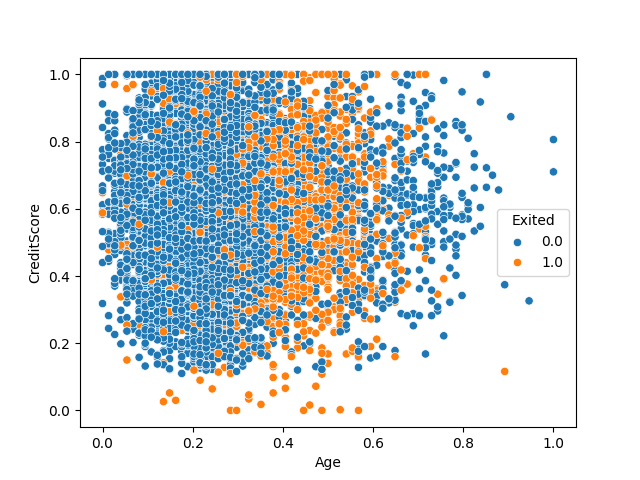
The age group of 30-40 have more chances to stay.

The age group of 40-50 are more likely to exit.

The pairplots between Age, Balance, Estimated Salary and Credit Score showed some inter dependence.

Graphs to showcase these statement is:

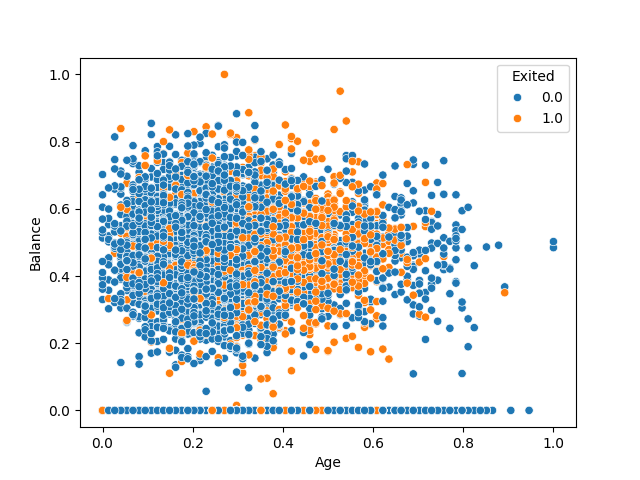
* + Age – Credit Score :



Observations:

In the scatter plot we can observe that there is observably more density of Exited points in Age range of 40-60 for all the values of Credit Scores.

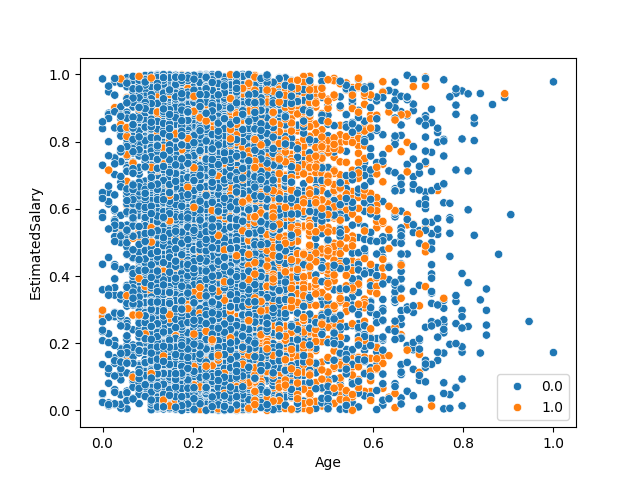
* + Age – Balance PairPlot:



Observations:

We can observe there is a density of points of Exited points in the Age range of 45 - 60 and with Balance range of 100000 – 150000

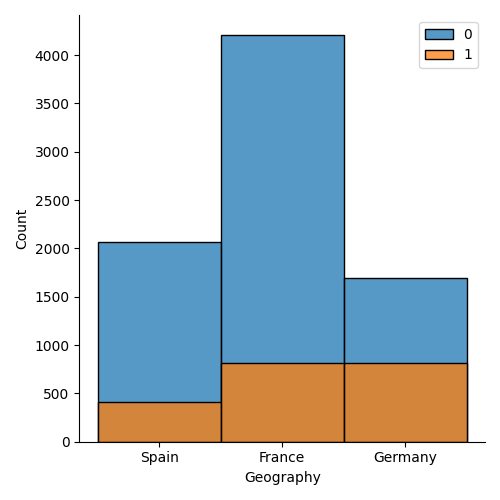
* + Age – Estimated Salary:



Observations:

We can observe there is density of people exited in the plot along the whole range of Estimated Salary, There is increased density of people exited in the rang of 40 - 60.

* Geography:



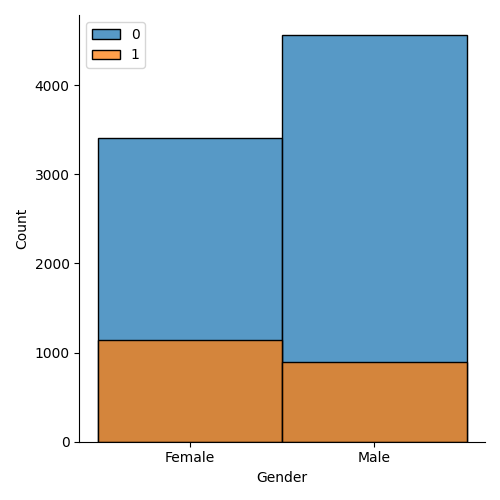
Observations:

Key observations made on the histogram plot is:

i) We can observe that the percentage of people from Germany Exiting the bank is almost 50% of that of people stayed.

ii) The percentage of people from spain and France who exited is very less compared to that of people who stayed.

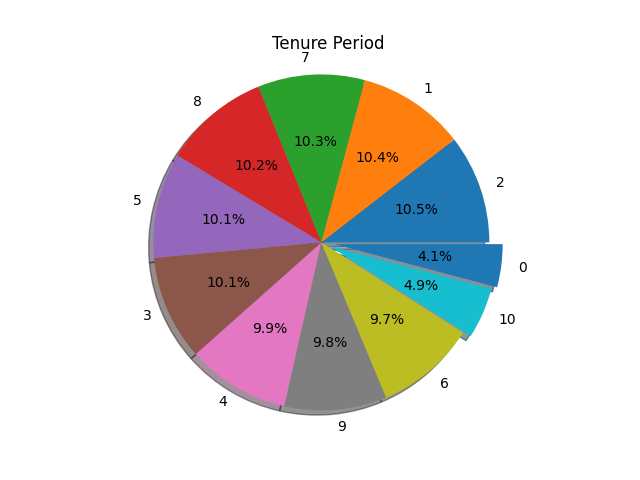
* Gender



Observations:

we can observe that females are more likely to exit the bank compared to male.

* Tenure:

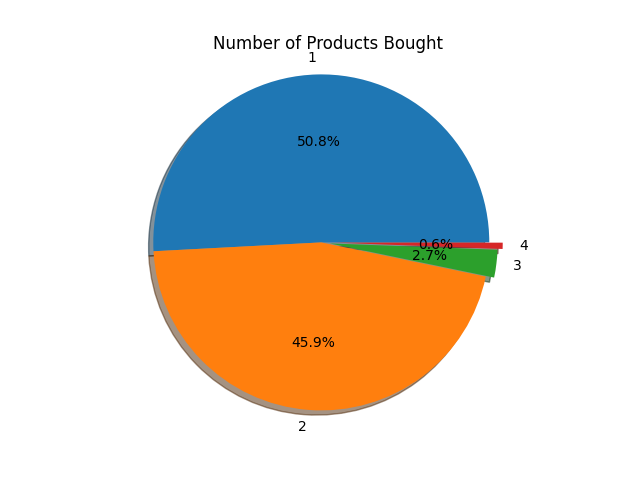


:

Observations:

We can observe that the count of people with tenures of 1-9 is almost same, and there are very less people with tenure period of 0 and the people with tenure period 10 is also tends to be very less.

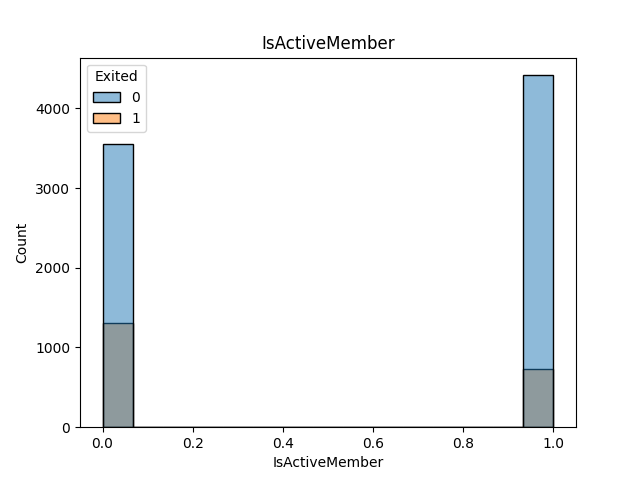
* NumOfProducts:



Observations:

We can observe more customers tends to have 1 or 2 products. And very less members have 3,4 products. Here the meaning of products is financial products like credit card , home loan , personal loam, insurance.

* IsActiveMember:



Observations:

We can observe that there are less people who exited who were a active member.

**DATA SPLITTING:**

The dataset is divided into X\_train,X\_test,y\_train,y\_test using the module train\_test\_split from sklearn.model\_selection library.

The size of test dataset is 20% of original dataset.

**MODEL TRAINING:**

The X\_train, X\_test,y\_train,y\_test are used to to train several ML Models and test datasets are used to calculated different Metrics to know the best model for classifying.

Table of Performance Metrics for Training Model:

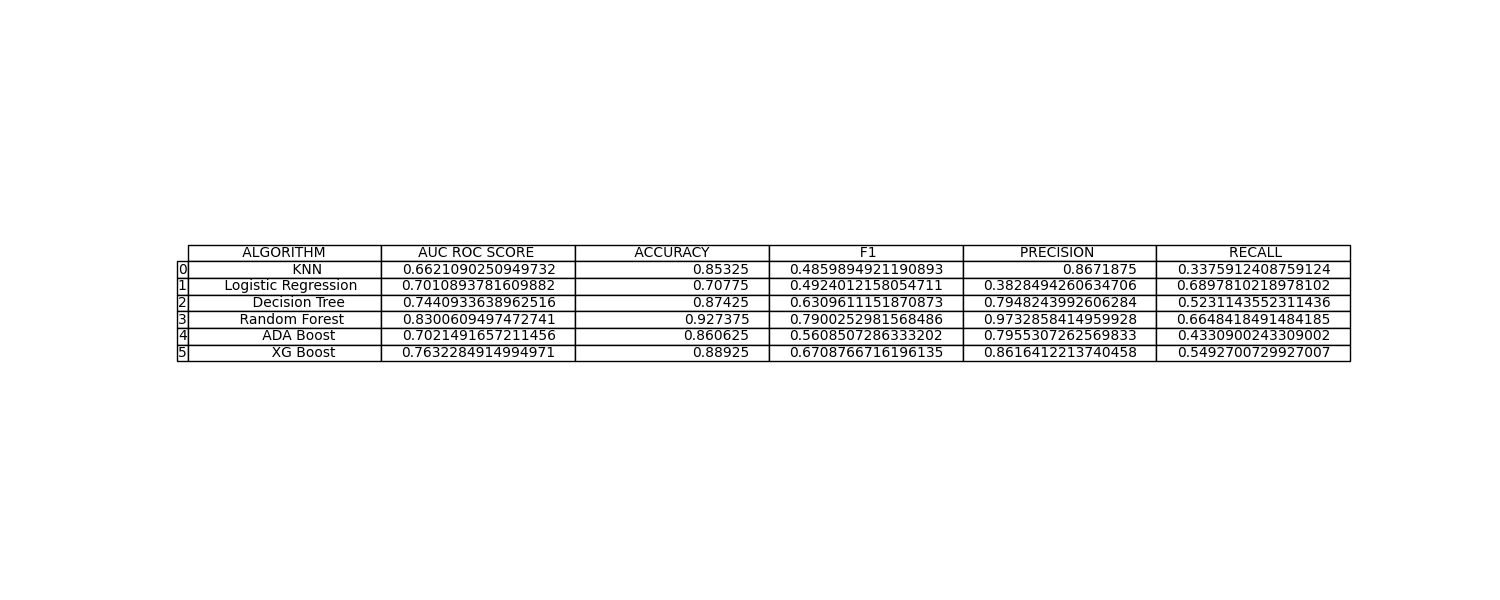
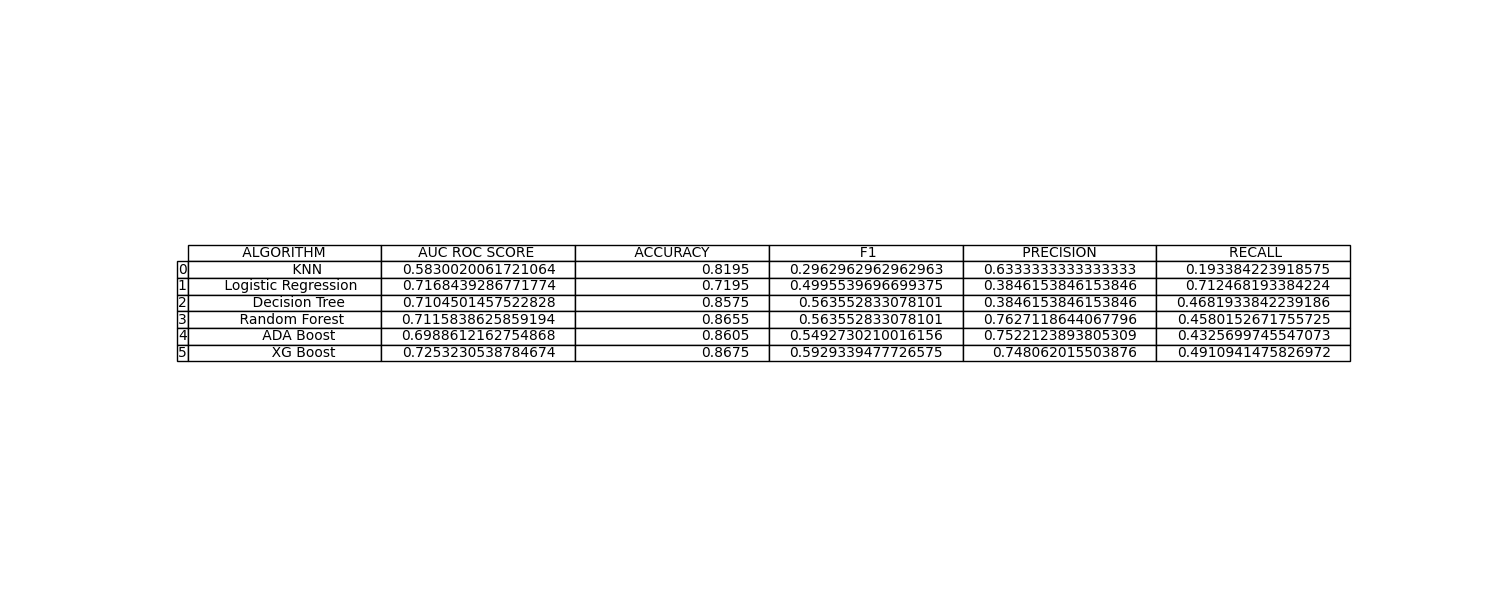


Table of Performance Metrics for Testing Model:



Observations:

The Algorithm Model of Random Forest is showing the best outcome results, metrics results for both training and testing datasets.

The ML Model of Random Forests gives best performance metrics values hence we will be using the Random Forest Model for our classification problem.

**RECOMMENDATIONS:**

1. The firm is required to implement new policies or features targeting the age category of 40 – 60 as they are the one’s who are more likely to leave the firm.
2. The firm should make sure the customers are well connected with the policies , and new things the firm offering ensuring they are active with the usage of their account, we observed the members who are not active members are more likely to leave the firm.
3. we can observe the percentage of females leaving the firm is more compared to percentage of males leaving the firm, so the firm should implement new policies catering to special needs of females.
4. The firm should try insisting more customers to get a Credit Card , getting a credit card helps the customers gain credit score , These will increase the number of products the customers posses.