|  |  |
| --- | --- |
|  | DATA MINING |
|  |  |
|  | VARUN KUMAR  PGP-DSBA  1/23/22 |

Contents

[A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage. 5](#_Toc93832628)

[1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 5](#_Toc93832629)

[IMPORT DATASET & Loading the necessary Libraries 5](#_Toc93832630)

[CHECKING THE SHAPE & MISSING VALUES IN THE DATA SET 5](#_Toc93832631)

[INFORMATION OF THE DATASET: 5](#_Toc93832632)

[DUPLICATES 5](#_Toc93832633)

[Summary of the dataframe: 6](#_Toc93832634)

[Univariate / Bivariate analysis 6](#_Toc93832635)

[SKEWNESS: 14](#_Toc93832636)

[MULTIVARIVARIATE ANALYSIS: 15](#_Toc93832637)

[1.2 Do you think scaling is necessary for clustering in this case? Justify 17](#_Toc93832638)

[1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. 18](#_Toc93832639)

[Cutting the Dendrogram with suitable clusters 18](#_Toc93832640)

[1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters. 19](#_Toc93832641)

[CLUSTER OUTPUT FOR ALL OBSERVATIONS 20](#_Toc93832642)

[Calculating the silhouette scores and silhouette width 21](#_Toc93832643)

[Observation 21](#_Toc93832644)

[1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 22](#_Toc93832645)

[HIERARCHIAL CLUSTER PROFILING: 22](#_Toc93832646)

[K\_MEANS CLUSTERING PROFILE: 23](#_Toc93832647)

[PROMOTIONAL STRATEGY 24](#_Toc93832648)

[An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets. 25](#_Toc93832649)

[2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 26](#_Toc93832650)

[Importing the Dataset and loading the necessary libraries 26](#_Toc93832651)

[SHAPE: 26](#_Toc93832652)

[INFORMATION: 26](#_Toc93832653)

[MISSING VALUES 27](#_Toc93832654)

[SUMMARY: 27](#_Toc93832655)

[DUPLICATES: 27](#_Toc93832656)

[Shape After Removing Duplicates: 27](#_Toc93832657)

[UNIVARIATE ANALYSIS & BIVARIATE ANALYSIS: 28](#_Toc93832658)

[CATEGORICAL VARIABLES: 33](#_Toc93832659)

[MULTIVARIATE ANALYSIS 38](#_Toc93832660)

[2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network. 42](#_Toc93832661)

[Splitting the data into Train and Test set 43](#_Toc93832662)

[Decision Tree Classifier: 43](#_Toc93832663)

[Random Forest Classifier 45](#_Toc93832664)

[Neural Network Classifier: 46](#_Toc93832665)

[2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model. 46](#_Toc93832666)

[CART MODEL PERFORMANCE EVALUATION: 46](#_Toc93832667)

[RF MODEL PERFORMANCE EVALUATION: 50](#_Toc93832668)

[ARTIFICIAL NEURAL NETWORK MODEL PERFORMANCE EVALUATION: 53](#_Toc93832669)

[2.4 Final Model: Compare all the models and write an inference which model is best/optimized. 57](#_Toc93832670)

[Classification Report Comparison for all the three models: 57](#_Toc93832671)

[ROC Curve for all the 3 models (Training data): 57](#_Toc93832672)

[ROC Curve for all the 3 models (Testing data): 58](#_Toc93832673)

[2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations. 59](#_Toc93832674)

[Figure 1 Description of Spending 6](#_Toc93832303)

[Figure 2 Histogram of Spending 7](#_Toc93832304)

[Figure 3 Boxplot of Spending 7](#_Toc93832305)

[Figure 4 Description of Advance Payments 8](#_Toc93832306)

[Figure 5 Histogram of Advance Payments 8](#_Toc93832307)

[Figure 6 Boxplot of Advanced Payment 8](#_Toc93832308)

[Figure 7 Probability of Full Payment Description 9](#_Toc93832309)

[Figure 8 Histogram of Probability of Full Payment 9](#_Toc93832310)

[Figure 9 Boxplot of Probability of Full Payment 9](#_Toc93832311)

[Figure 10 Description of Current Balance 10](#_Toc93832312)

[Figure 11 Histogram of Current Balance 10](#_Toc93832313)

[Figure 12 Boxplot of Current Balance 10](#_Toc93832314)

[Figure 13 Description of Credit Limit 11](#_Toc93832315)

[Figure 14 Histogram of Credit Limit 11](#_Toc93832316)

[Figure 15 Boxplot of Credit Limit 11](#_Toc93832317)

[Figure 16 Description of Minimum Payment 12](#_Toc93832318)

[Figure 17 Histogram of Minimum Payment 12](#_Toc93832319)

[Figure 18 Boxplot of Minimum Payment 12](#_Toc93832320)

[Figure 19 Description of Max Spent in Single Shopping 13](#_Toc93832321)

[Figure 20 Histogram of Max Spent in Single Shopping 13](#_Toc93832322)

[Figure 21 Boxplot of Max Spent in Single Shopping 13](#_Toc93832323)

[Figure 22 Pair plot of Bank Dataset 15](#_Toc93832324)

[Figure 23 Correlation Matrix of Bank Dataset 16](#_Toc93832325)

[Figure 24 Heatmap of Bank Dataset 16](#_Toc93832326)

[Figure 25 Scaled Bank Dataset 17](#_Toc93832327)

[Figure 26 Hierarchical Clustering 18](#_Toc93832328)

[Figure 27 Hierarchical Clustering (Truncate mode) 18](#_Toc93832329)

[Figure 28 Bank Dataset with Clusters 19](#_Toc93832330)

[Figure 29 The sum distance within the centroids 19](#_Toc93832331)

[Figure 30 Cluster output for all the Observations 20](#_Toc93832332)

[Figure 31 K-elbow Plot 20](#_Toc93832333)

[Figure 32 Bank dataset with Sil\_width & clusters 21](#_Toc93832334)

[Figure 33 Dataset 26](#_Toc93832335)

[Figure 34 Dataset Info 26](#_Toc93832336)

[Figure 35 Summary of Dataset 27](#_Toc93832337)

[Figure 36 Description of Age 28](#_Toc93832338)

[Figure 37 Histogram of Age 28](#_Toc93832339)

[Figure 38 Boxplot of Age 29](#_Toc93832340)

[Figure 39 Description of Commission 29](#_Toc93832341)

[Figure 40 Histogram of Commission 30](#_Toc93832342)

[Figure 41 Boxplot of Commission 30](#_Toc93832343)

[Figure 42 Description of Duration 31](#_Toc93832344)

[Figure 43 Histogram of Duration 31](#_Toc93832345)

[Figure 44 Boxplot of Duration 31](#_Toc93832346)

[Figure 45 description of Sales 32](#_Toc93832347)

[Figure 46 Histogram of Sales 32](#_Toc93832348)

[Figure 47 Boxplot of Sales 33](#_Toc93832349)

[Figure 48 Bar Plot of Agency Code 33](#_Toc93832350)

[Figure 49 Boxplot of Agency Code 34](#_Toc93832351)

[Figure 50 Boxplot of Type 34](#_Toc93832352)

[Figure 51 Bar Plot of Type 35](#_Toc93832353)

[Figure 52 Boxplot of Channel used 35](#_Toc93832354)

[Figure 53 Bar Plot of Channel Used 36](#_Toc93832355)

[Figure 54 Bar Plot of Product 36](#_Toc93832356)

[Figure 55 Boxplot of Product 37](#_Toc93832357)

[Figure 56 Boxplot of Destination 37](#_Toc93832358)

[Figure 57 Count Plot of destination 38](#_Toc93832359)

[Figure 58 Pair plot 39](#_Toc93832360)

[Figure 59 Heatmap 40](#_Toc93832361)

[Figure 60 Outliers 41](#_Toc93832362)

[Figure 61 After treating outliers 41](#_Toc93832363)

[Figure 62 Decision Tree 44](#_Toc93832364)

[Figure 63 Variable Importance Decision Tree 45](#_Toc93832365)

[Figure 64 Variable Importance RF 46](#_Toc93832366)

[Figure 65 Decision Tree ROC curve for Training Data 47](#_Toc93832367)

[Figure 66 Decision Tree ROC curve for Test Data 48](#_Toc93832368)

[Figure 67 ROC curve for training data RF 51](#_Toc93832369)

[Figure 68 ROC curve for test data RF 53](#_Toc93832370)

[Figure 69 ANN ROC Curve for training data 55](#_Toc93832371)

[Figure 70 ROC curve ANN for Test data 56](#_Toc93832372)

[Figure 71 ROC of training data for all three models 58](#_Toc93832373)

[Figure 72 ROC curve for testing data of all Three Models 58](#_Toc93832374)

[Table 1 Dataset 5](#_Toc93832375)

[Table 2 Summary 6](#_Toc93832376)

[Table 3 Cluster Frequency Table 22](#_Toc93832377)

[Table 4 Summary of Clusters & Features 22](#_Toc93832378)

[Table 5 Cluster Frequency Table 23](#_Toc93832379)

[Table 6 Summary Of k\_means clusters & Features 23](#_Toc93832380)

[Table 7 Decision Tree Confusion Matrix for Training Data 47](#_Toc93832381)

[Table 8 Decision Tree Confusion Matrix for Test Data 49](#_Toc93832382)

[Table 9 Confusion matrix for train data RF 50](#_Toc93832383)

[Table 10 Confusion Matrix For test data RF 52](#_Toc93832384)

[Table 11 Confusion Matrix ANN for training data 54](#_Toc93832385)

[Table 12 Confusion Matrix For ANN Test data 55](#_Toc93832386)

[Table 13 Classification Report comparison 57](#_Toc93832387)

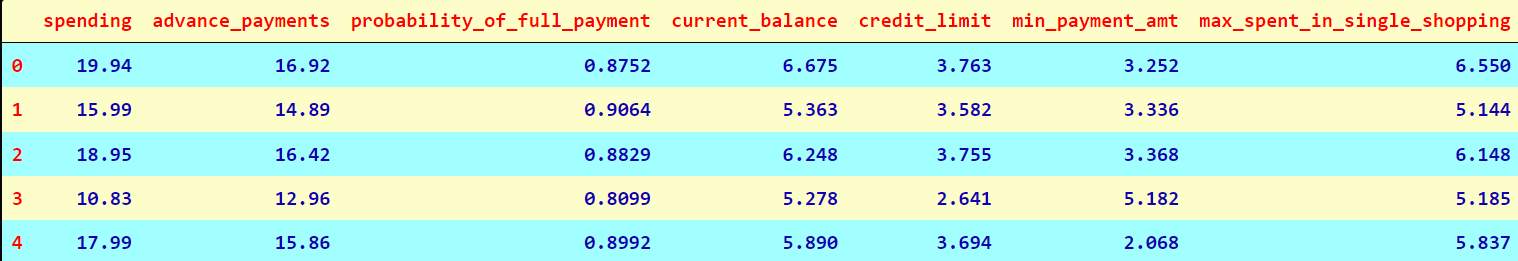
**Problem 1: Clustering**

# A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

## **1.1** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

### IMPORT DATASET & Loading the necessary Libraries

Table Dataset



### CHECKING THE SHAPE & MISSING VALUES IN THE DATA SET

• The Shape of the Data is (210,7)

• There were No Null Values in the Data

### INFORMATION OF THE DATASET:

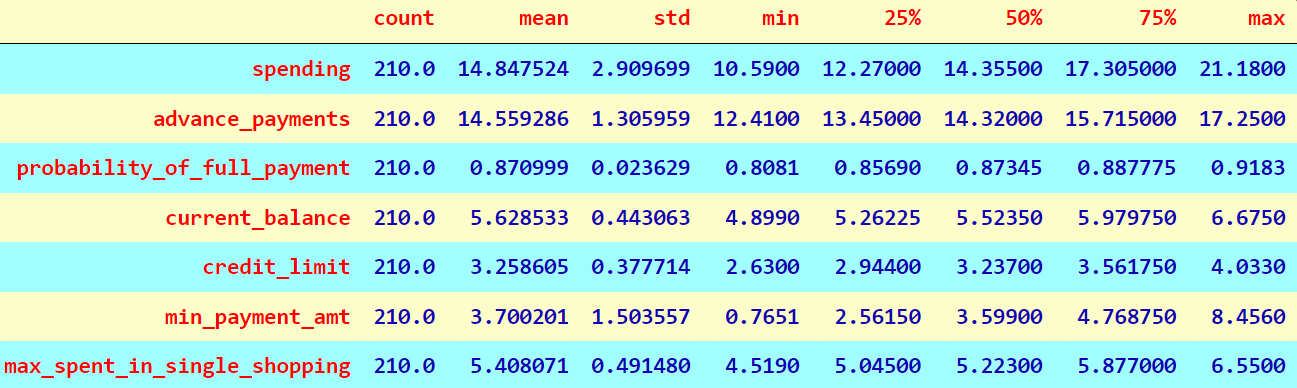
• From the data it was observed all the values are Float Values.

DUPLICATES:

There are no duplicate rows present in the dataset.

### Summary of the dataframe:

Table Summary



From the above summary, it can be seen that the Std Deviation of spending is high when compared with the other variables. Also, there are no null values present in any of the variables.

### Univariate / Bivariate analysis

• Boxplot is an indication of how the values in the data are spread out and also indicates if any outlier is present.

• Histogram gives the univariant set of observations.

Helps us to understand the distribution of data in the dataset. With univariate analysis we can find patterns and we can summarize the data and have understanding about the data to solve our business problem.

#### Spending

Figure Description of Spending

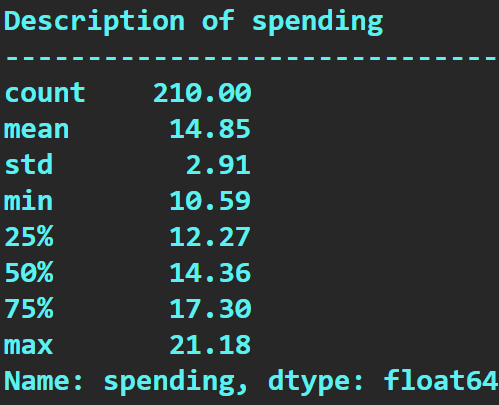
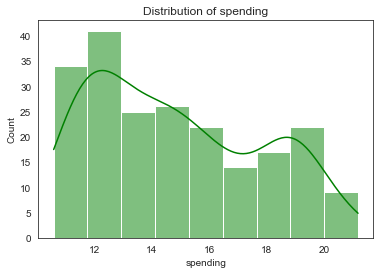
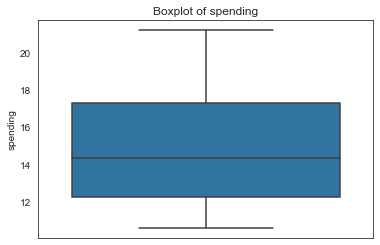


Figure Histogram of Spending



The histogram shows the distribution of data from 10 to 22.

Figure Boxplot of Spending



The box plot of the spending variable shows no outliers.

#### Advanced Payments

Figure Description of Advance Payments

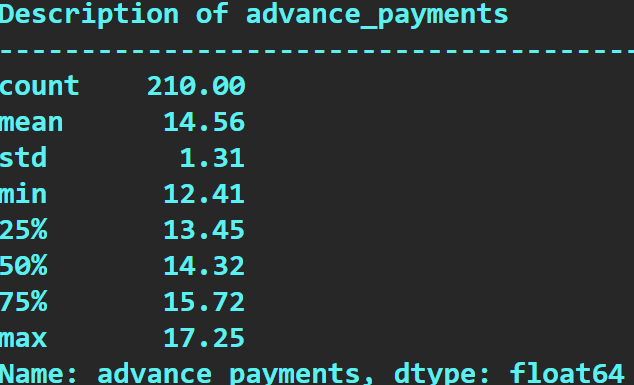
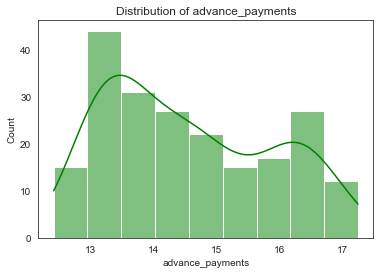
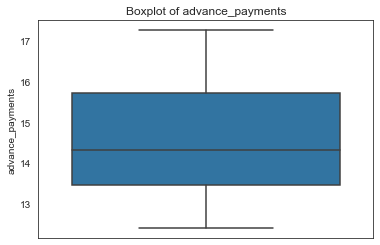


Figure Histogram of Advance Payments



The histogram shows the distribution of data from 12 to 17.

Figure Boxplot of Advanced Payment



The box plot of the advance payment variable shows no outliers.

#### Probability of Full Payment

Figure Probability of Full Payment Description

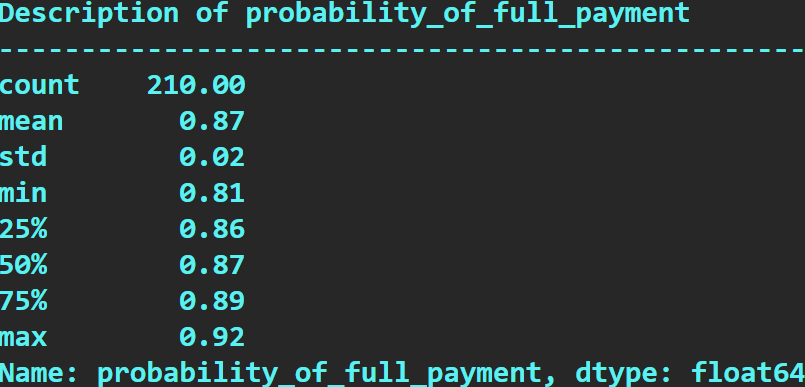
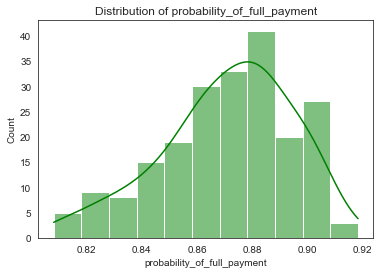
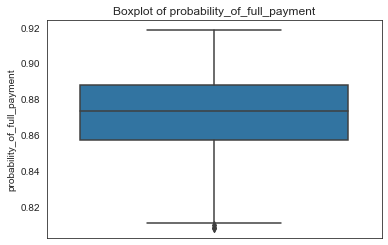


Figure Histogram of Probability of Full Payment



The histogram shows the distribution of data from 0.80 to 0.92.

Figure Boxplot of Probability of Full Payment



The box plot of the probability of full payment variable shows few outliers.

#### Current Balance

Figure Description of Current Balance

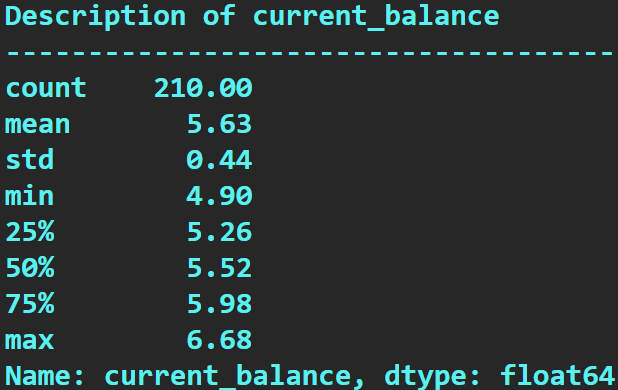
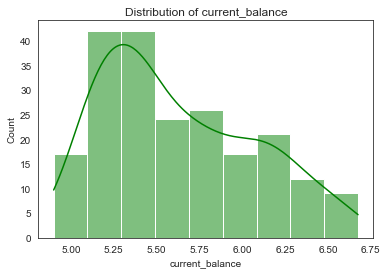
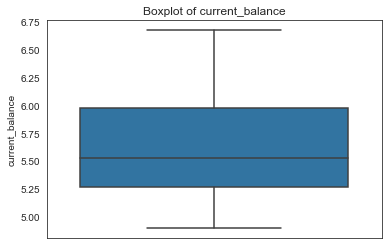


Figure Histogram of Current Balance



The histogram shows the distribution of data from 5.0 to 6.5.

Figure Boxplot of Current Balance



The box plot of the current balance variable shows no outliers.

#### Credit Limit

Figure Description of Credit Limit

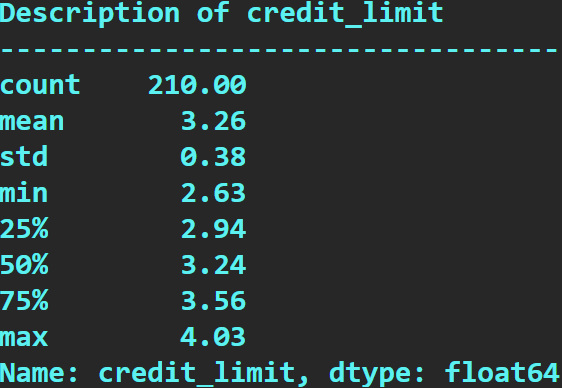
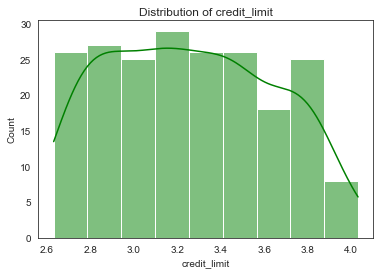
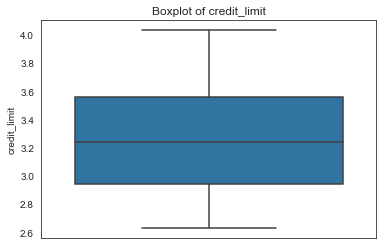


Figure Histogram of Credit Limit



The histogram shows the distribution of data from 2.5 to 4.0.

Figure Boxplot of Credit Limit



The box plot of the credit limit variable shows no outliers.

#### Minimum Payment Amount

Figure Description of Minimum Payment

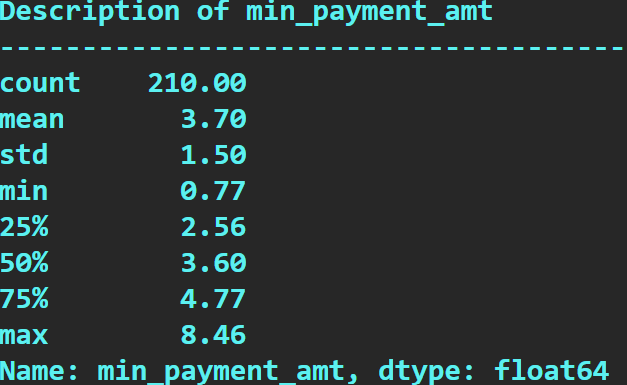
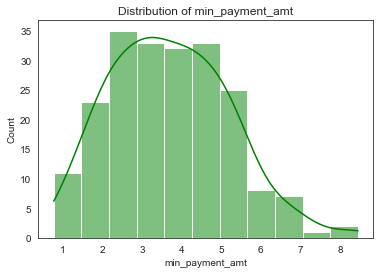
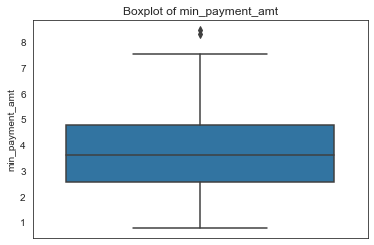


Figure Histogram of Minimum Payment



The histogram shows the distribution of data from 0.7 to 8.

Figure Boxplot of Minimum Payment



The box plot of the min payment amount variable shows few outliers.

#### Maximum Spent in Single Shopping

Figure Description of Max Spent in Single Shopping

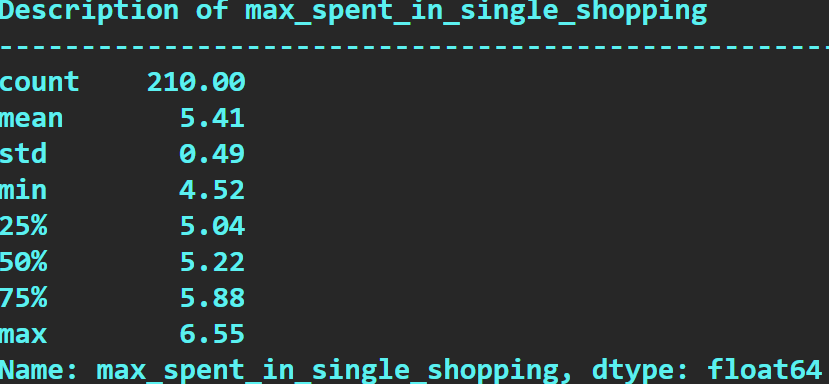
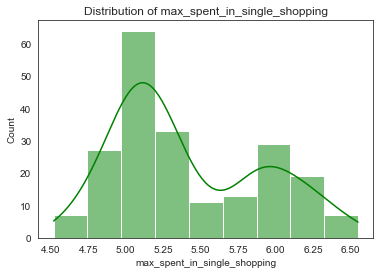
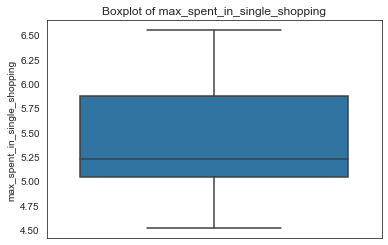


Figure Histogram of Max Spent in Single Shopping



The histogram shows the distribution of data from 4.5 to 6.5.

Figure Boxplot of Max Spent in Single Shopping



The box plot of the max spent in single shopping variable shows no outliers.

From, the above Univariate Analysis, the data was summarised & patterns can be visualised. Outlier is present only in one variable which can be seen in min\_payment\_amt which clearly indicates that there are only a few customers whose minimum payment amount falls on the higher side on an average and probability\_of\_full\_payment has outliers on the lower side which indicates there are few customers whose probability to pay to the bank in full is on the lower side of average. min\_payment\_amt & probability\_of\_full\_payment variables alone have a very small Outliers, hence the requirement for Outlier Treatment is not necessary.

### SKEWNESS:

Skewness of Spending is **0.39702715402072153**

Skewness of Advance Payment is **0.38380604212562563**

Skewness of Probability of Full Payment is **-0.5341035521949097**

Skewness of Current Balance is **0.5217206481959239**

Skewness of Credit Limit is **0.13341648969738146**

Skewness of Minimum Payment Amount is **0.3987925792256687**

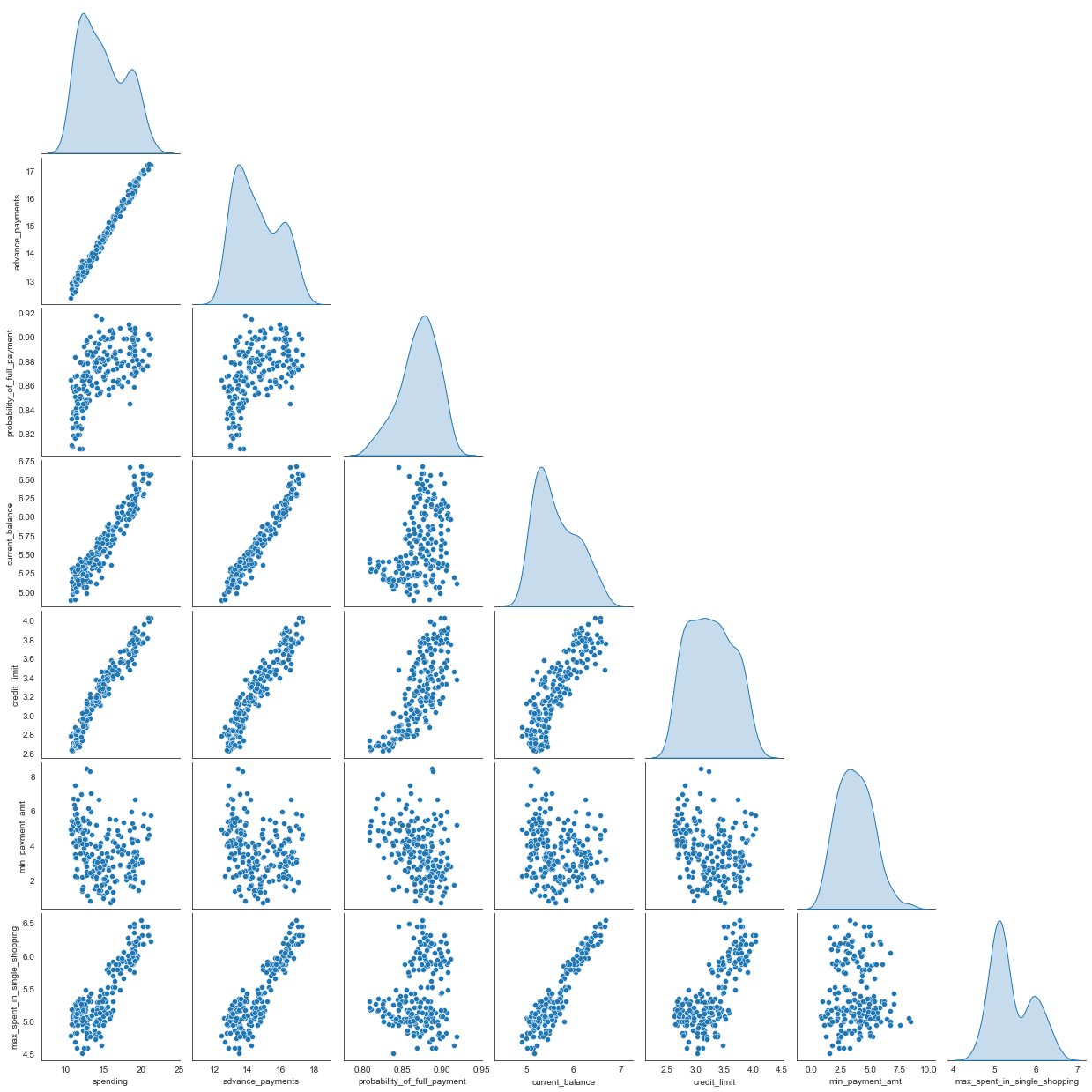
Skewness of Max Spent in Single Shopping is **0.5578758322317957**

Probability of \_full\_payment is **-0.5341035521949097** (negatively skewed). From, the Histogram Visualization the data distribution takes on from 0.80 to 0.92 which is a good factor. Only this Variable is negatively skewed apart from other variables which are positively skewed.

### MULTIVARIVARIATE ANALYSIS:

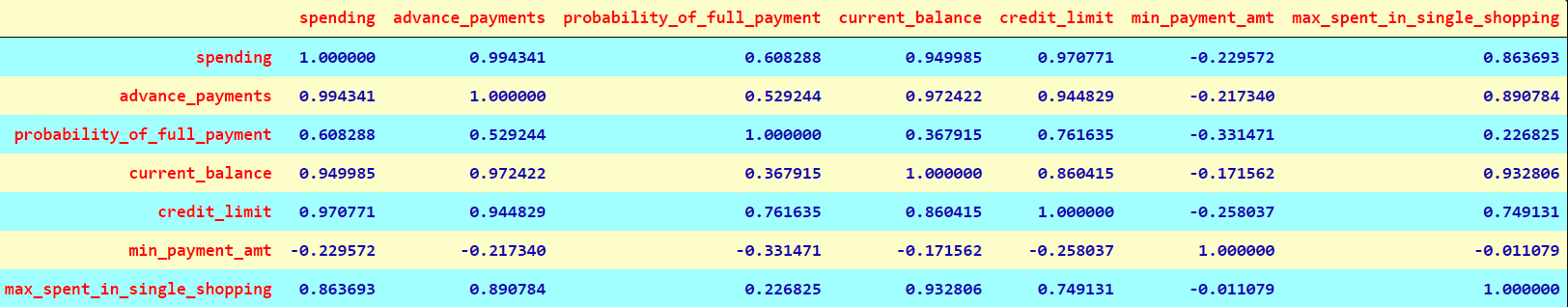
#### Checking for Multi Collinearity

Figure Pair plot of Bank Dataset



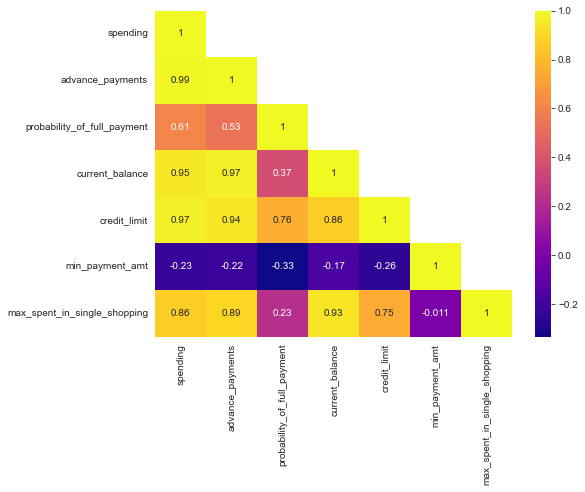
#### Correlation

Figure Correlation Matrix of Bank Dataset



#### Heatmap

Figure Heatmap of Bank Dataset



From the above Heatmap Visualization it can be seen that there is Strong Positive Correlation between Spending & advance\_payments. Spending & current\_balance is highly correlated. Spending & credit\_limit are highly correlated. Advance\_payment and current\_balance is highly correlated, Advance\_payament and credit\_limit is highly correlated. Advance\_payment and max\_spent\_in\_single\_shopping is highly correlated current\_balance and max\_spent\_in\_single\_shopping is highly correlated.

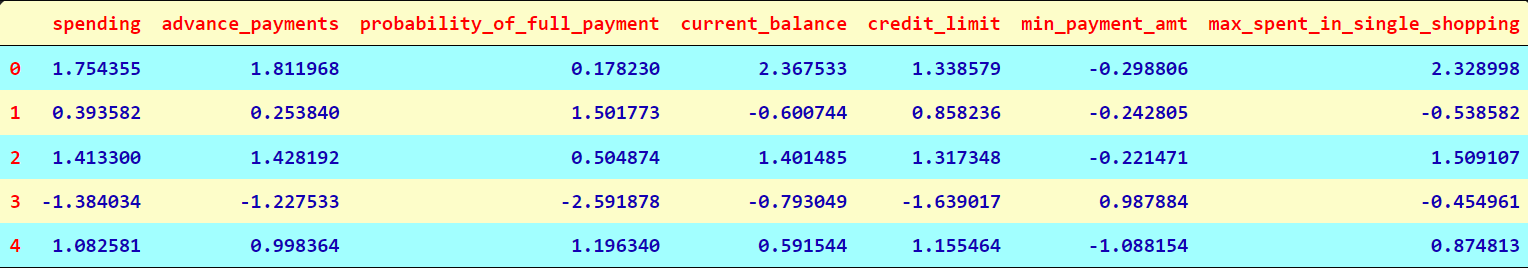
We can conclude based on the data that those who have high credit limit & those who have high Current Balance amount spends more. Min\_payment\_amt is not correlated with any of the variables which won’t influence changes in spending, current balance or credit limit. Probability of full payments are much higher for those customers who have a higher credit limit.

## **1.2** Do you think scaling is necessary for clustering in this case? Justify

Yes, Scaling needs to be done as the values of the variables are in different scales. (i.e., Values of the variables ranging from 100’s to 1000’s up to 10000’s). Spending, advance payments are in different values and this may get more weightage. In order to perform Analysis all of these different variables need to be converted to one scale. As we can check from the summary of the data, there is a significance difference between the standard deviation of spending (2.909), advance\_payments (1.305), min\_payment\_amt (1.503) as compared to the other features. By performing Scaling, all of the values will fall in the same relative range. I have used standard scalar for scaling.

Below is the snapshot of scaled data.

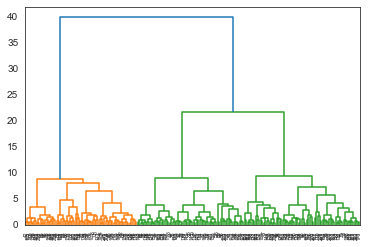
Figure Scaled Bank Dataset



## **1.3** Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

I have used Hierarchical clustering method to create optimum clusters & splitting the dataset based upon these clusters. Hierarchical clustering relies using clustering techniques to find a hierarchy of clusters, where this hierarchy resembles a dendrogram.

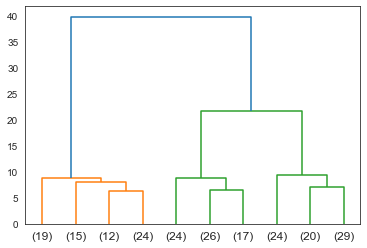
Figure Hierarchical Clustering



The above dendrogram indicates that all the data points have been clustered into different clusters by Ward’s Method. To find the optimum cluster to solve the problem further we can use truncate mode = lastp. Where, we can give last p = 10, according to industry set base value.

### Cutting the Dendrogram with suitable clusters

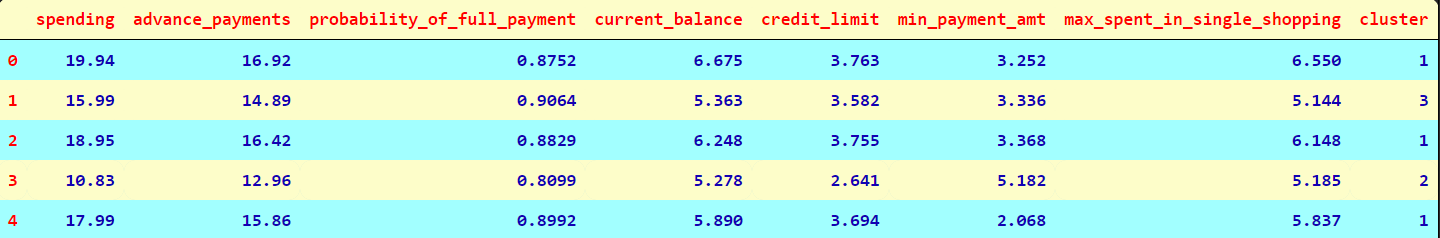
Figure Hierarchical Clustering (Truncate mode)



Now, by using the above method we can see that the data has been clustered into 3 different Clusters. We can also visualize that the maximum customers fall into the green Cluster.

Next step is to append these clusters into the dataset. We can use fcluster with criterion = 'maxclust' where a cut is defined based on the number of clusters.

Figure Bank Dataset with Clusters

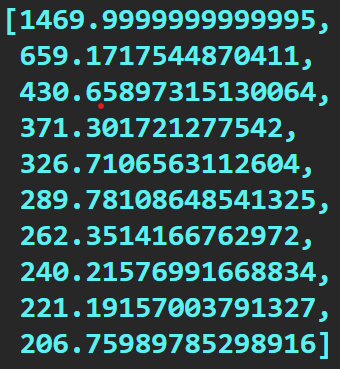


## **1.4** Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

In K-Means, each cluster is associated with a centroid. K-means is a centroid-based algorithm where we calculate the distances to assign a point to a cluster.

We, apply K-Means Technique to the scaled dataset & identify the clusters formed. Then we will calculate the value of inertia and store it in WSS.

Figure The sum distance within the centroids

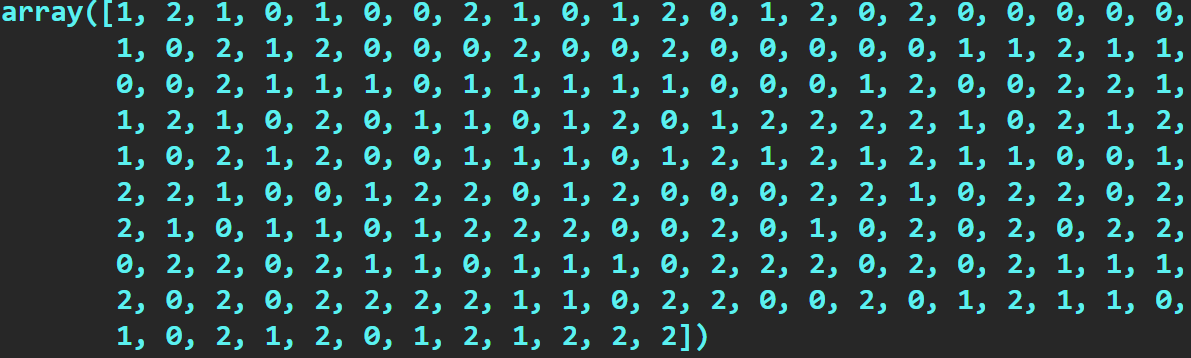


The above gives a clear idea on the inertia values of Clusters from 1 to 11.

Randomly we decide to give n\_clusters = 3 and we look at the distribution of clusters according to the n\_clusters.

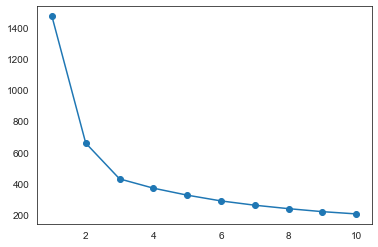
### CLUSTER OUTPUT FOR ALL OBSERVATIONS

Figure Cluster output for all the Observations



The two methods to determine the optimal number of clusters are within sum of squares (wss) method and average silhouette scores method.

Figure K-elbow Plot



From the above visualisation the optimal number of clusters to be taken for k-means clustering is 3 since as per the elbow it can be easily seen in the curve that after 3 the curve gets flat.

### Calculating the silhouette scores and silhouette width

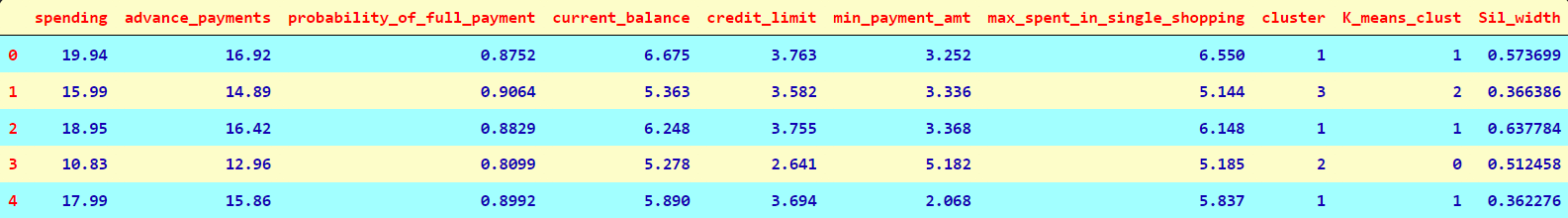
silhouette\_score is **0.40072705527512986**

*If minimum individual score, i.e., silhouette width is positive, then all the scores are positive and our process is good to go.*

silhouette\_width is **0.002713089347678376**

The Silhouette Score for 3 Clusters is better than nearest possible Optimum Clusters which is n\_clusters = 4, n\_clusters=5.Also the elbow curve seen above shows, there is no huge drop in values so we can finalise on 3 clusters. Hence, as per wss method and silhouette score we can conclude that the optimal number of clusters is 3. The 3 Group Cluster Solution gives a pattern based on each category of spending.

Figure Bank dataset with Sil\_width & clusters

****

This table shows the clusters to the dataset and also individual sil\_width score.

### Observation

By K- Mean’s method we can at cluster 3 we find it optimal after there is no huge drop in inertia values. Also, the elbow curve seems to show similar results. The silhouette width score of the K – means also seems to very less value that indicates all the data points are properly clustered to the cluster. There is no mismatch in the data points with regards to clustering Cluster grouping based on the dendrogram, 3 or 4 looks good. Did the further analysis, and based on the dataset had gone for 3 group cluster and three group cluster solution gives a pattern based on high/medium/low spending with max\_spent\_in\_single\_shopping (high value item) and probability\_of\_full\_payment (payment made).

## **1.5** Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

### HIERARCHIAL CLUSTER PROFILING:

#### Cluster Group Profiles

Group 1: High Spending; Group 2: Low Spending; Group 3: Medium Spending

#### Cluster Frequency

Table Cluster Frequency Table

|  |  |  |
| --- | --- | --- |
| Cluster Group Number | Spending | Frequency |
| 1 | High Spending | 70 |
| 2 | Low Spending | 67 |
| 3 | Medium Spending | 73 |

Table Summary of Clusters & Features



For Cluster 1, Spending is highest, averaging 18371 which is highest among three clusters. Highest advance payments around 1614 which is highest among three clusters. Probability of Full Payment is very high, averaging around 0.8844 which is highest among three clusters.

For Cluster 2, The average Spending of this cluster is on lower side, averaging 11872. Highest advance payments around 1325 which is lowest among the three clusters. Probability of Full Payment is the least amongst other clusters, averaging around 0.848. Current Balance is around 5238 which is least among three clusters. Credit Limit is least for this cluster ranging around 28485. min\_payment\_amt is 494 which is max in this cluster. max\_spent\_in\_single\_shopping is around 5122.

For Cluster 3, Average Spending of this cluster is 14199. Highest advance payments around 1423 Probability of Full Payment are on the higher side, averaging around 0.879. Current Balance is around 5478 which is average among three clusters. Credit Limit is around 32264 which is average among three clusters. min\_payment\_amt is 261 which is least among three cluster. Max\_spent\_in\_single\_shopping is the least around 5086.

### K\_MEANS CLUSTERING PROFILE:

#### Cluster Group Profiles

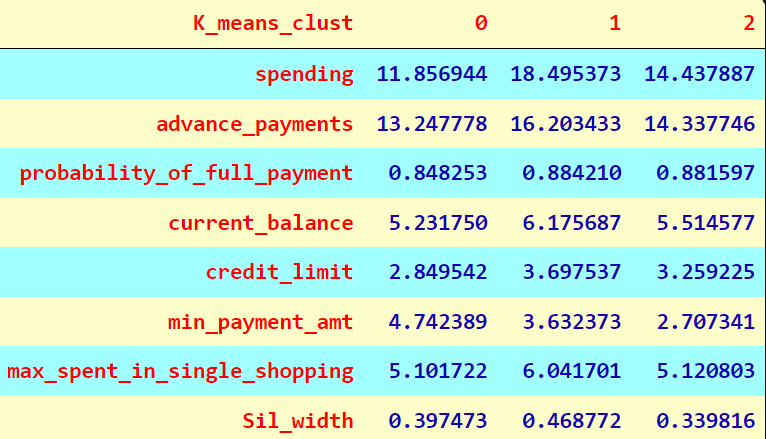
Group 0: Low Spending; Group 1: High Spending; Group 2: Medium Spending

#### Cluster Frequency

Table Cluster Frequency Table

|  |  |  |
| --- | --- | --- |
| Cluster Group Number | Spending | Frequency |
| 0 | Low Spending | 72 |
| 1 | High Spending | 67 |
| 2 | Medium Spending | 71 |

Table Summary Of k\_means clusters & Features



### PROMOTIONAL STRATEGY

HIERARCHIAL Clusters:

Cluster 1 are premium customers. Bank should focus on Cluster 1 as the customers in this cluster have higher spending, highest advance payments, highest Current Balance and highest credit limit. This segment appears to be upper class and can be targeted using various offers such as cards with rewards and loyalty points for every spent. Increase their credit limit & giving any reward points might increase their purchases. Maximum max\_spent\_in\_single\_shopping is high for this group, so can be offered discount/offer on next transactions or upon full payment. Give loan against the credit card, as they are customers with good repayment record. Tie up with luxury brands, which will drive more one\_time\_maximun spending.

Cluster 2 are low spending customers; Poor spending customers has the least Credit limit and so maybe they spend least and also they have least current balance. Bank can also think of providing them offers for shopping at various websites & promotions that may increase their max\_spent\_in\_single\_shopping by tying up with basic amenities groups such as groceries, utilities etc. Customers should be given remainders for payments often. This Cluster need huge promotions schemes & targets must be set in order to move customers from this cluster to medium spending Cluster.

Cluster 3 medium spending customers. Bank should give customers in this Cluster more promotional offers because there are more chances that these customers may move to Cluster 1 of high spending. Increase spending habits by tying with ecommerce sites, premium hotels, luxury etc. Hence Bank should provide more promotional offers for customers & encourage them to spend more.

For K-means Clusters:

Cluster 0 are low spending customers; Poor spending customers has the least Credit limit and so maybe they spend least and also they have least current balance. Bank can also think of providing them offers for shopping at various websites & promotions that may increase their max\_spent\_in\_single\_shopping by tying up with basic amenities groups such as groceries, utilities etc. Customers should be given remainders for payments often. This Cluster need huge promotions schemes & targets must be set in order to move customers from this cluster to medium spending Cluster.

Cluster 1 are premium customers. Bank should focus on Cluster 1 as the customers in this cluster have higher spending, highest advance payments, highest Current Balance and highest credit limit. This segment appears to be upper class and can be targeted using various offers such as cards with rewards and loyalty purchases.

Cluster 2 medium spending customers. Bank should give customers in this Cluster more promotional offers because there are more chances that these customers may move high spending category of customers. Increase spending habits by tying with ecommerce sites, premium hotels, luxury etc. Hence Bank should provide more promotional offers for customers & encourage them to spend more.

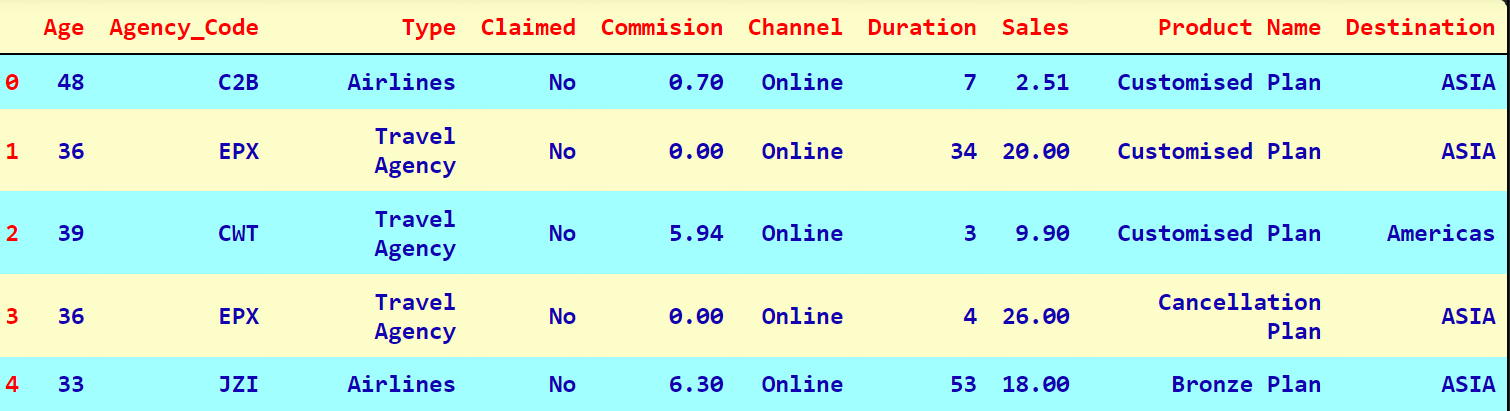
PROBLEM 2 : CART-RF-ANN

# An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

## **2.1** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

### Importing the Dataset and loading the necessary libraries

Figure Dataset



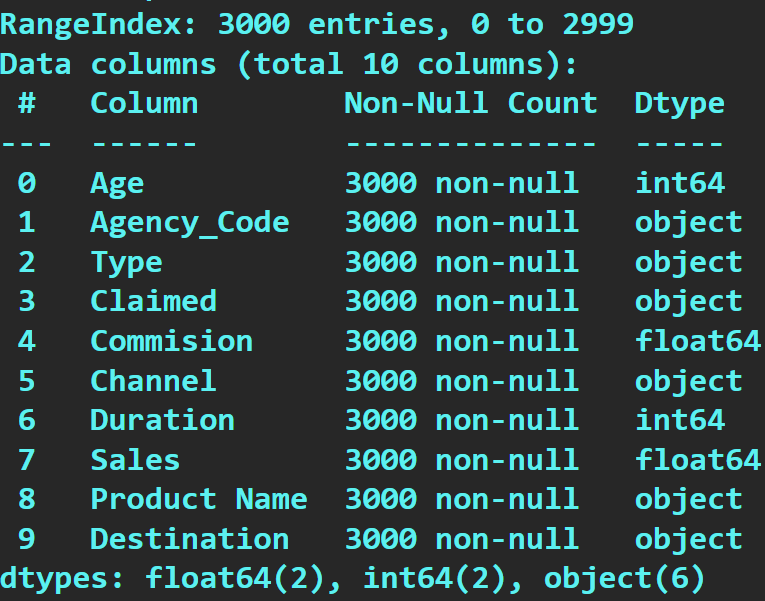
### SHAPE:

The shape of the dataset is (3000,10).

### INFORMATION:

The Info of the dataset shows that the dataset contains integer, float & object values. Henceforth, we need to convert the object datatypes into a numeric value.

Figure Dataset Info

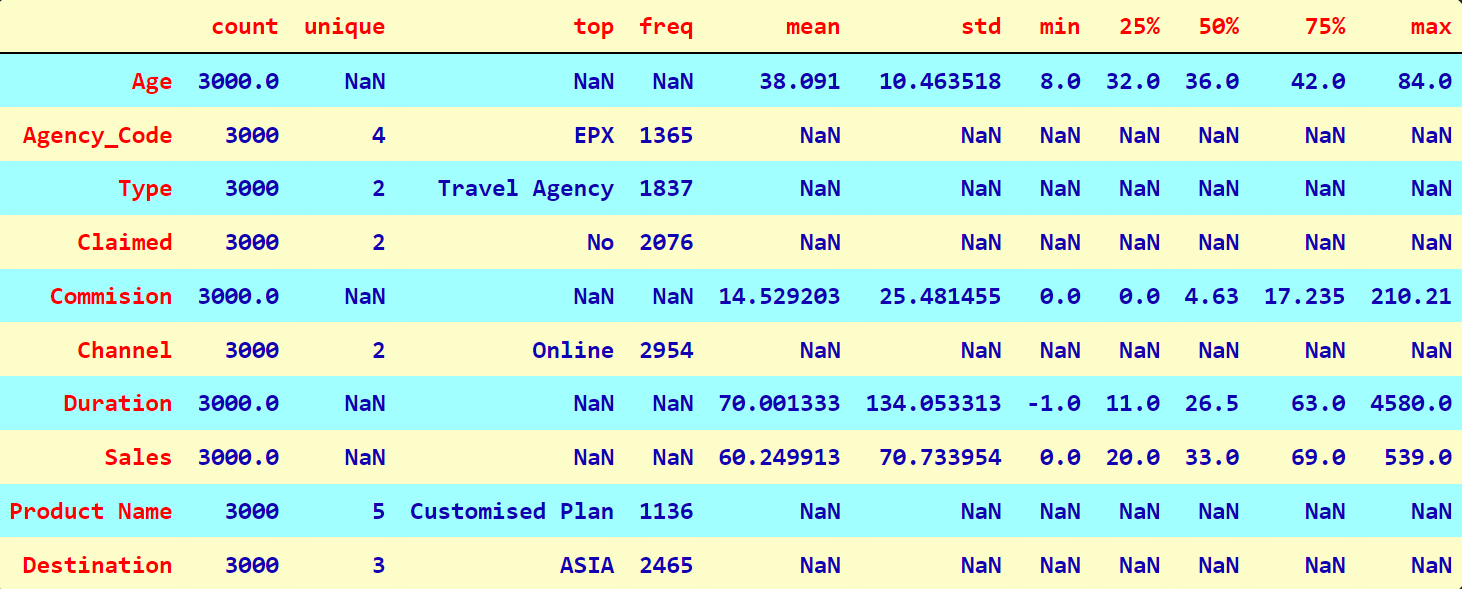


MISSING VALUES:

There are no missing values in the dataset.

### SUMMARY:

Figure Summary of Dataset



There are totally 4 numeric variables & 6 Categorical Variables. The Most Preferred type is Travel Agency; Channel is online; Customised Plans are most preferred by the customers. Destination ASIA is the preferred destination by the customers. There are only 4 continuous variables Age, Commission, Duration and Sales, the result is shown for them only.

### DUPLICATES:

Number of duplicate rows = 139. We need to remove these duplicate rows from the dataset.

### Shape After Removing Duplicates:

The shape of the dataset is (2861,10).

### UNIVARIATE ANALYSIS & BIVARIATE ANALYSIS:

#### Age:

Figure Description of Age

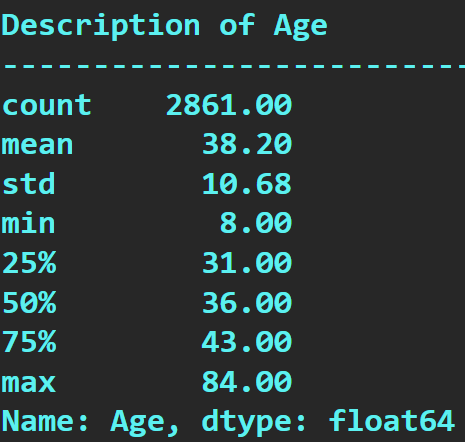


Figure Histogram of Age

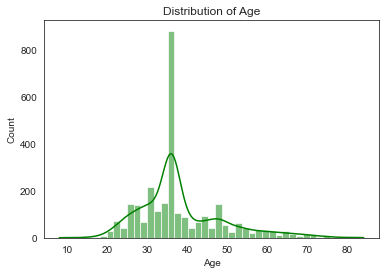
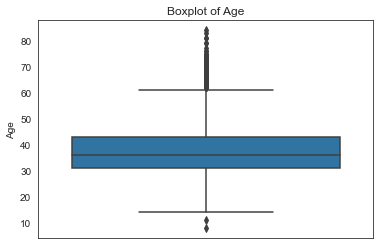


Figure Boxplot of Age



Age has outliers and its positively skewed **1.1025661500650201**

#### Commission:

Figure Description of Commission

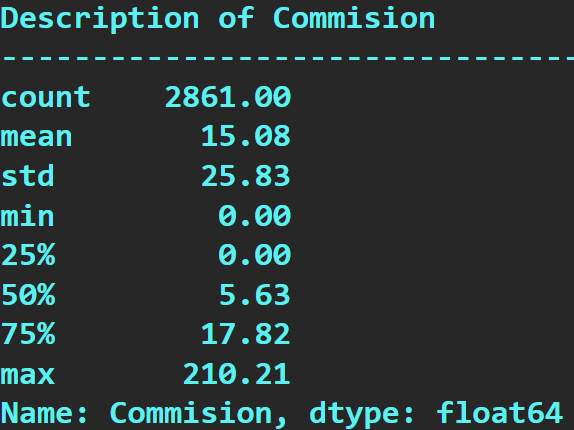


Figure Histogram of Commission

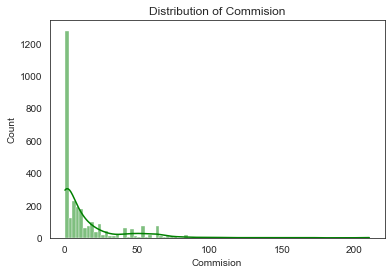
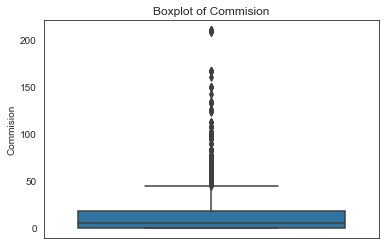


Figure Boxplot of Commission



Commission has outliers and its positively skewed **3.1031126292410716**

#### Duration:

Figure Description of Duration

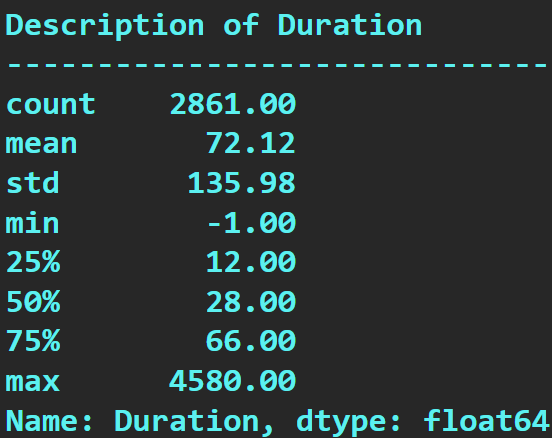


Figure Histogram of Duration

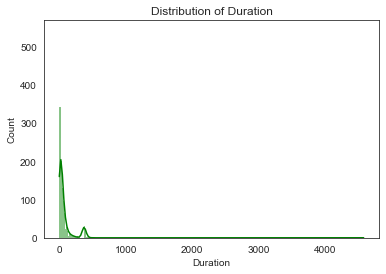
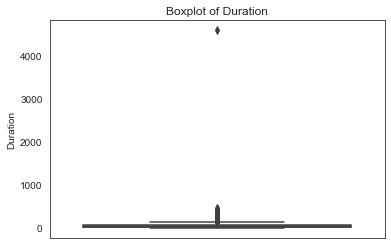


Figure Boxplot of Duration



Duration has outliers and its positively skewed **13.778867077621834**

#### Sales:

Figure description of Sales

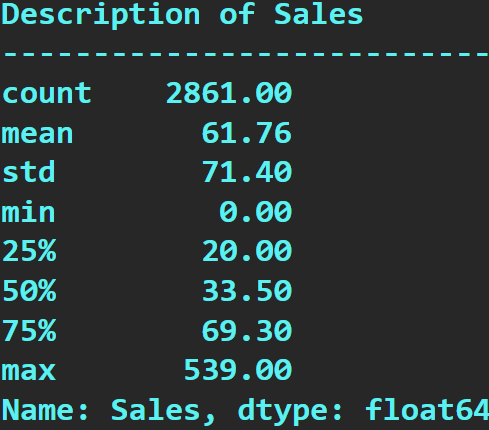


Figure Histogram of Sales

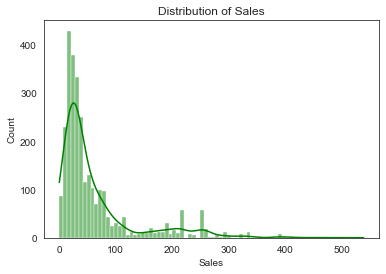
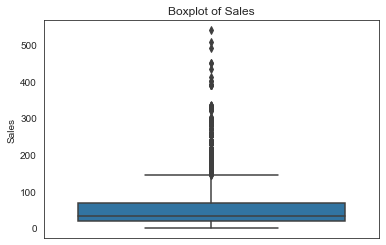


Figure Boxplot of Sales



Sales has outliers and its positively skewed **2.3434132352067008**

### CATEGORICAL VARIABLES:

#### Agency Code:

Figure Bar Plot of Agency Code

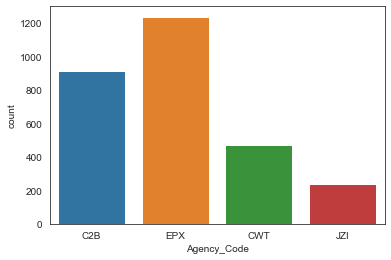
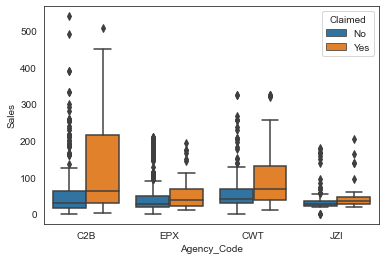


Figure Boxplot of Agency Code



From the above, we can see that Agency EPX has higher frequency and agency C2B has highest number of claims.

#### Type

Figure Boxplot of Type

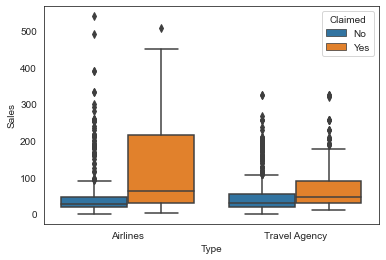
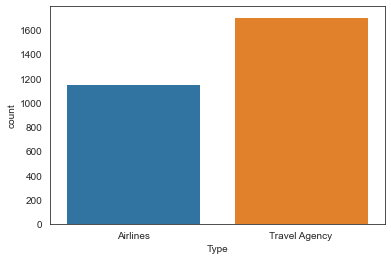


Figure Bar Plot of Type



Travel Agencies have been more Sales over Airline. Insurance claim in Airlines, has been on higher side comparatively to Travel Agency which had higher Sale frequencies than Airlines.

#### Channel

Figure Boxplot of Channel used

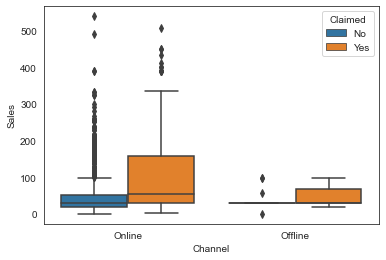
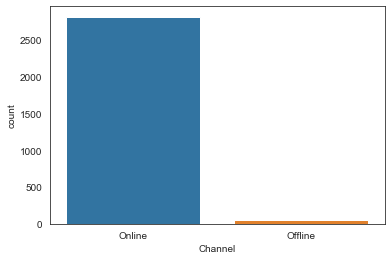


Figure Bar Plot of Channel Used



Channel Online is very much used than offline. Insurance Claim has been more in Online.

#### Product:

Figure Bar Plot of Product

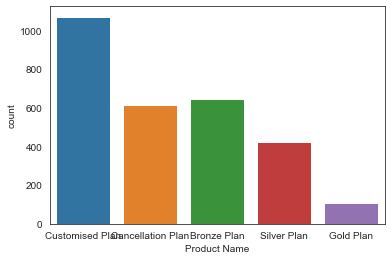
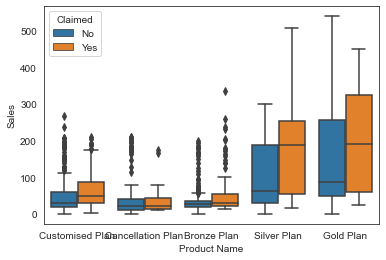


Figure Boxplot of Product



Customised products are more in demand and gold plan members claimed the most.

#### Destination:

Figure Boxplot of Destination

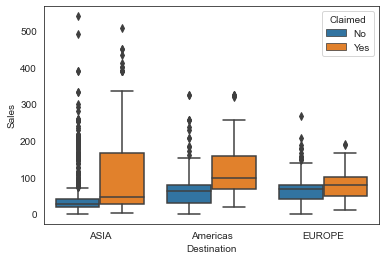
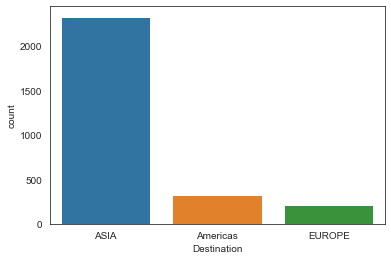


Figure Count Plot of destination



Most Sale & Most Claimed Destination is ASIA and people travel ASIA claimed most.

Further to our inference, we can see that maximum of the customers doing a claim in our data belong to age group of 30-50 years and it is observed that age group between 30-40 contribute to the highest number of claims, the type of Tour Agency was Travel Agency, Product name was Customised Plan, Channel was Online and Destination was Asia.

### MULTIVARIATE ANALYSIS

Figure Pair plot

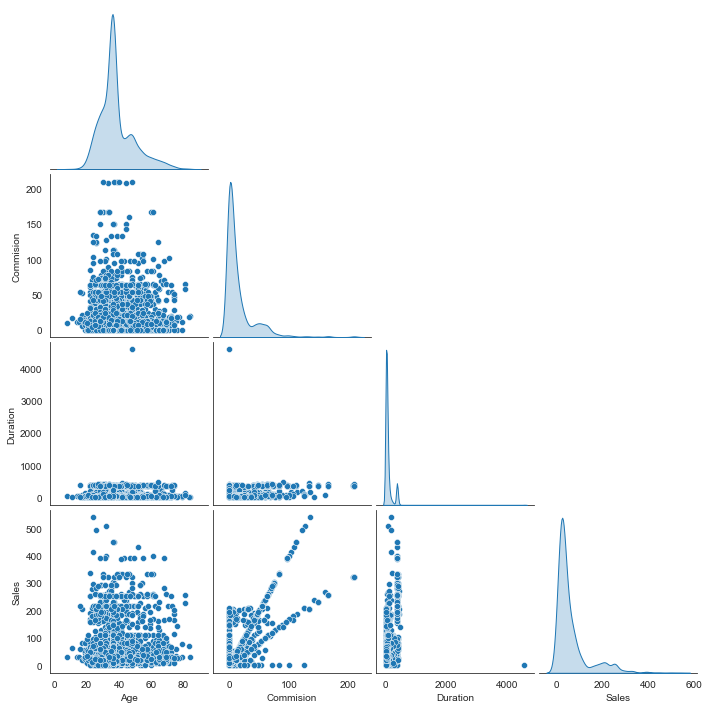
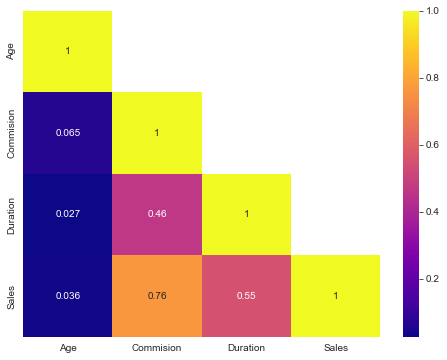


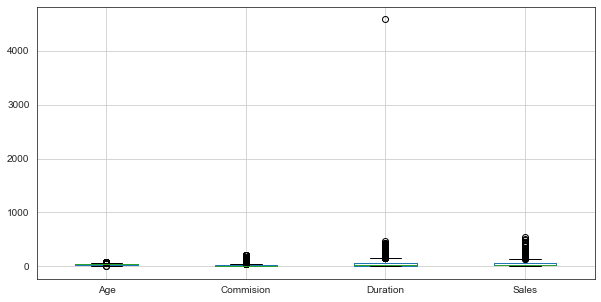
Figure Heatmap



From the above Heatmap, we can see there is no negative correlation & only positive correlation is observed. Not much multi-collinearity is observed.

#### CHECKING FOR OUTLIERS:

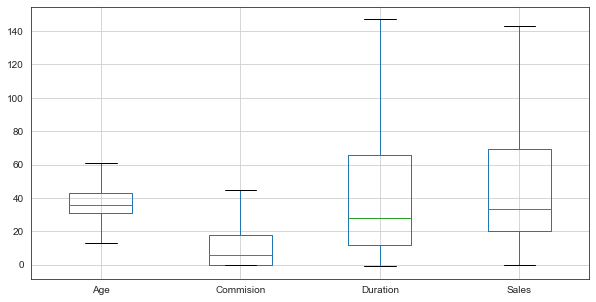
Figure Outliers



By checking the Boxplots for all the continuous variables, we can conclude that a very high number of outliers are present in all the continuous variables. Hence, it is better to remove the presence of outliers which lies in our dataset to further build the models for the problem.

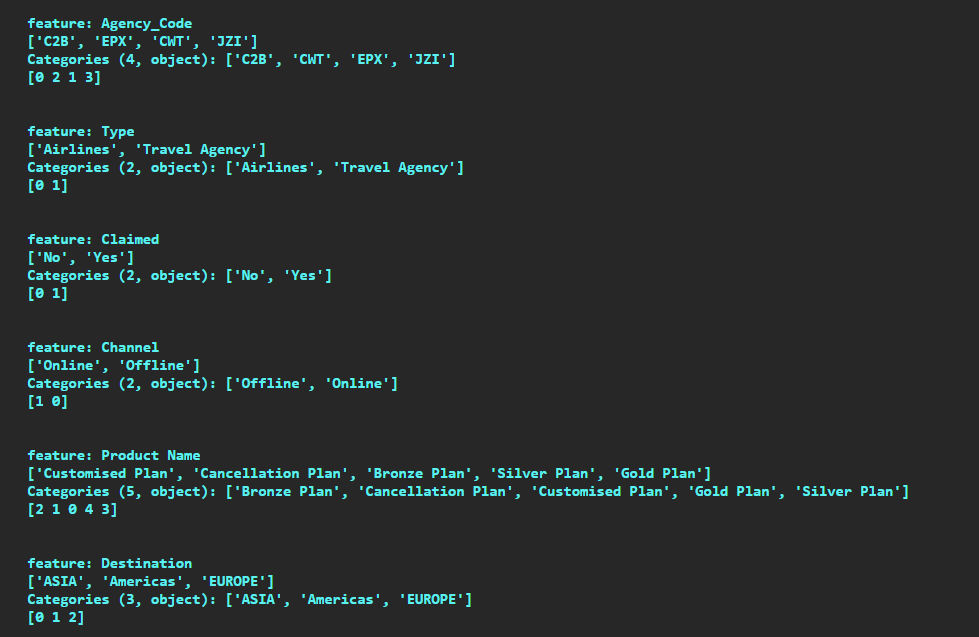
After treating outliers:

Figure After treating outliers

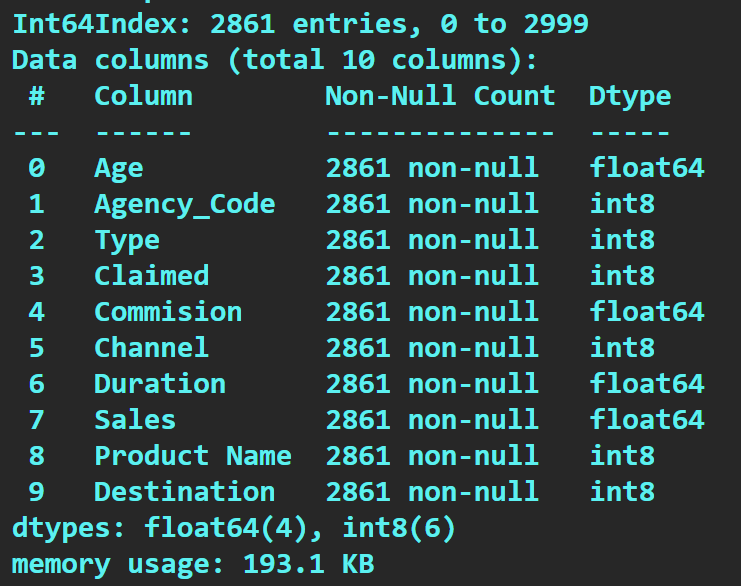


## **2.2** Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

The Models which we are going to build does not take in objects but only Numeric. Hence object data types are converted into Categorical variables.



Now, all the objects variables have been converted into Categorical ones.



Proportion of 1s and 0s in our target variable.

0 is **0.680531**

1 is **0.319469**

***There is no issue of class imbalance, here we have reasonable proportions in both the classes***

### Splitting the data into Train and Test set

Claimed is our Target Variable. Extraction of the target column into separate vectors for training set and test set is required. Now we can split our dataset.

Test size we have given as 0.30 as we want 30% of the data is to be tested. Random state we have given as 1234, as it will give similar results with similar random state respectively. The dimensions of the training and test data are below; the samples are almost equally distributed between the train and test datasets:

X\_train (2002, 9)

X\_test (859, 9)

train\_labels (2002,)

test\_labels (859,)

### Decision Tree Classifier:

#### With GINI CRITERION:

We used Grid Search. It is essentially an optimization algorithm which enables us to classify best parameters for the optimization of the problem from a list of parameters. Fitting our train dataset to the grid search. After fitting the values, we will get the best parameters using some inbuilt functions. Best parameters are following:

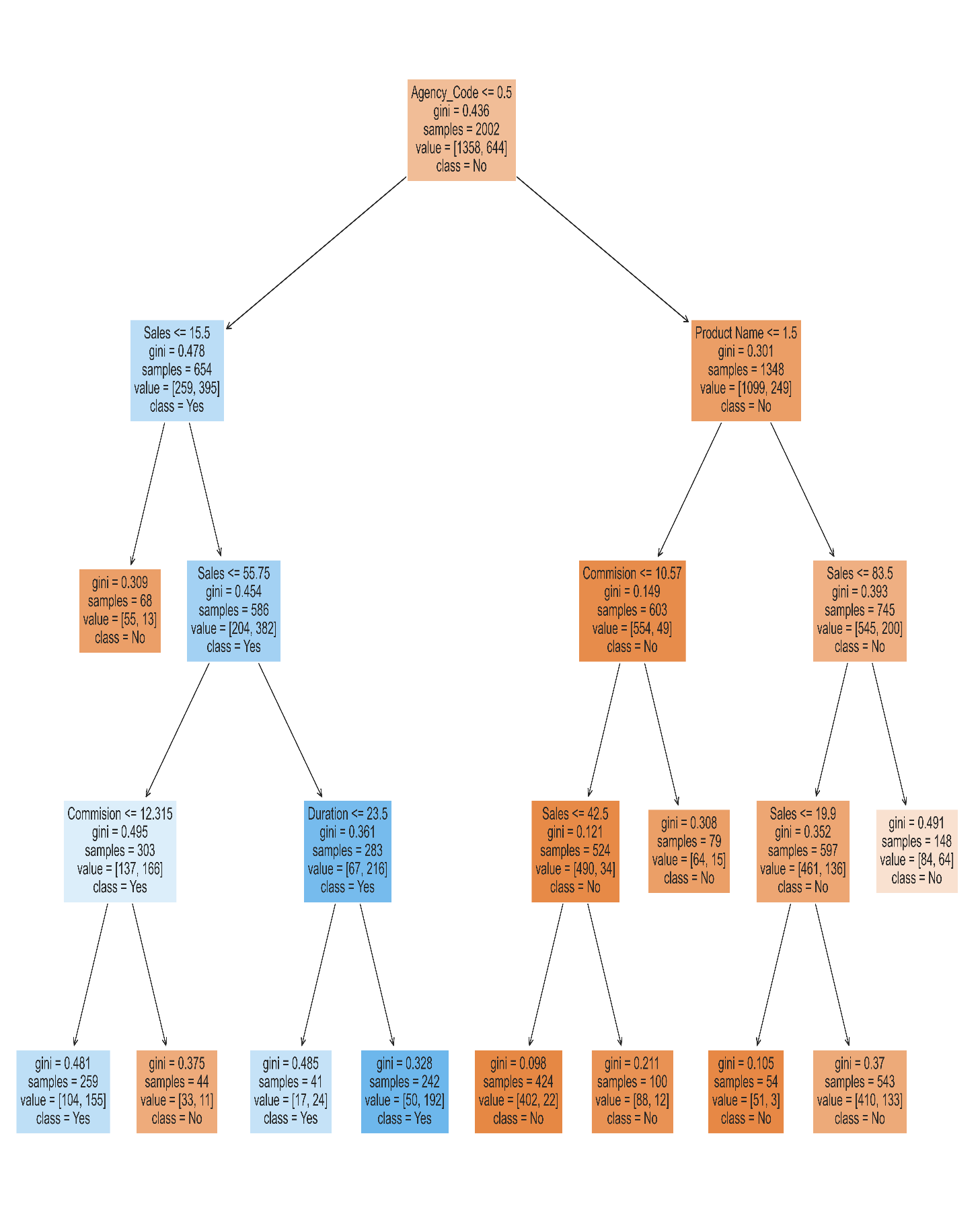
**max\_depth: 4,**

**min\_samples\_leaf: 40,**

**min\_samples\_split: 200**

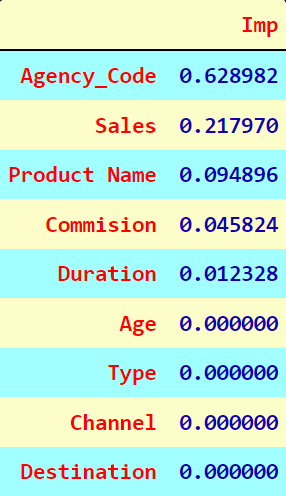
By adding the best grid parameters & saving it on to an output dot file & by pasting the code on http://webgraphviz.com/ to view the tree chart as follows: which will aid in carrying out the test & train data analysis. Below is a glimpse of a decision tree:

Figure Decision Tree

****

#### Variable Importance:

Figure Variable Importance Decision Tree



Random Forest Classifier:

The random forest classifier can use for both classification and the regression task. Random forest classifier will handle the missing values. When we have more trees in the forest, random forest classifier won’t over fit the model. We can model the random forest classifier for categorical values also. Now that, splitting of the data into Train and Test set is already done. RF Model can be built by getting the best parameters. For best parameters, once again we are going to use Grid Search followed by fit function. Best parameters are following:

**max\_depth: 6,**

**max\_features: 7,**

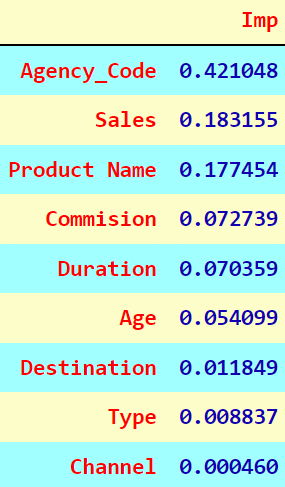
**min\_samples\_leaf: 8,**

**min\_samples\_split: 45,**

**n\_estimators: 200**

#### Variable Importance:

Figure Variable Importance RF



### Neural Network Classifier:

We get the best parameters & best grid as shown below:

**hidden\_layer\_sizes: 500,**

**max\_iter: 5000,**

**solver: 'adam',**

**tol: 0.01**

## **2.3** Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

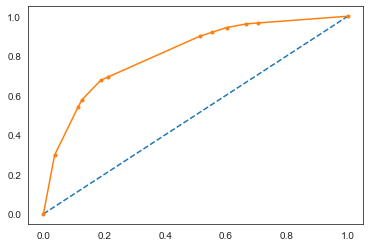
### CART MODEL PERFORMANCE EVALUATION:

Decision Trees are commonly used in data mining with the objective of creating a model that predicts the value of dependent variable based on the values of several independent variables.

AUC and ROC Plot for the training data To Predict the probabilities of train data.Calculate the AUC as well as the ROC curve and plotting them.

**AUC: 0.810**

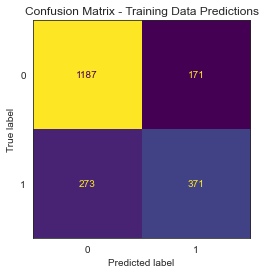
Figure Decision Tree ROC curve for Training Data



From the Graph it can be seen that we have derived the AUC Values as 0.810 & ROC Plot for the Train Data.

#### Confusion Matrix and Accuracy for the training data:

Table 7 Decision Tree Confusion Matrix for Training Data



Accuracy for training data is **0.7782217782217782.**

#### Classification Report for the training data:

**precision recall f1-score support**

**0 0.81 0.87 0.84 1358**

**1 0.68 0.58 0.63 644**

**accuracy 0.78 2002**

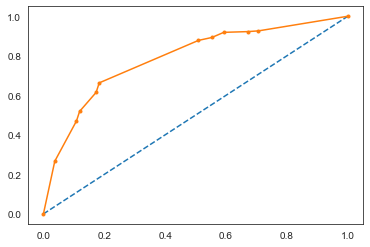
**macro avg 0.75 0.73 0.73 2002**

**weighted avg 0.77 0.78 0.77 2002**

#### AUC and ROC for the testing data:

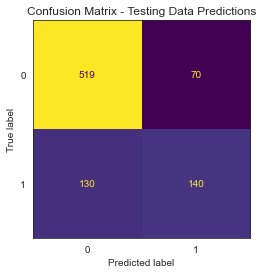
**AUC: 0.787**

Figure Decision Tree ROC curve for Test Data

****

#### Confusion Matrix and Accuracy for the testing data:

Table 8 Decision Tree Confusion Matrix for Test Data



Accuracy score for test data is **0.7671711292200233.**

#### Classification Report for the testing data:

**precision recall f1-score support**

**0 0.80 0.88 0.84 589**

**1 0.67 0.52 0.58 270**

**accuracy 0.77 859**

**macro avg 0.73 0.70 0.71 859**

**weighted avg 0.76 0.77 0.76 859**

Train Data Accuracy: 0.78; Test Data Accuracy: 0.77;

Train Data Precision: 0.68; Test Data Precision: 0.67;

Train Data Recall: 0.58; Test Data Recall: 0.52;

Train Data f1-score: 0.63; Test Data f1-score: 0.58;

Train Data AUC: 0.810; Test Data AUC: 0.787;

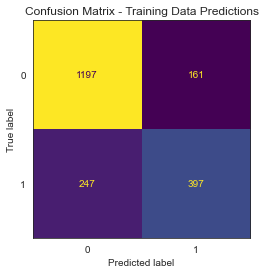
From observing the characteristics of the CART training & testing data set, we can observe that the results are close & almost similar. Overall, the model is a good model.

### RF MODEL PERFORMANCE EVALUATION:

Random forest classifier can handle the missing values. When there are more trees in the forest, random forest classifier won’t over fit the model. We Can model the random forest classifier for categorical values also. The model is built with dependant variable as Claimed, and considering all other variables as independent variables.

#### Confusion Matrix and Accuracy for the training data:

Table 9 Confusion matrix for train data RF



Accuracy for training data is **0.7962037962037962.**

#### Classification Report for the training data:

**precision recall f1-score support**

**0 0.83 0.88 0.85 1358**

**1 0.71 0.62 0.66 644**

**accuracy 0.80 2002**

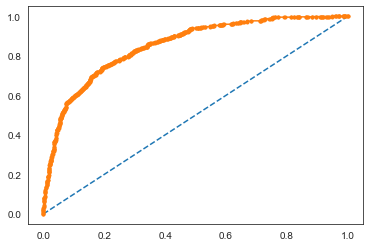
**macro avg 0.77 0.75 0.76 2002**

**weighted avg 0.79 0.80 0.79 2002**

#### AUC & ROC PLOT FOR TRAINING DATA:

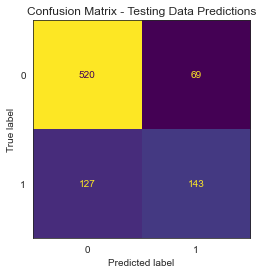
**AUC: 0.852**

Figure ROC curve for training data RF



#### Confusion Matrix and Accuracy for the testing data:

Table 10 Confusion Matrix For test data RF



Accuracy of test data is **0.7718277066356228.**

#### Classification Report for the testing data:

**precision recall f1-score support**

**0 0.80 0.88 0.84 589**

**1 0.67 0.53 0.59 270**

**accuracy 0.77 859**

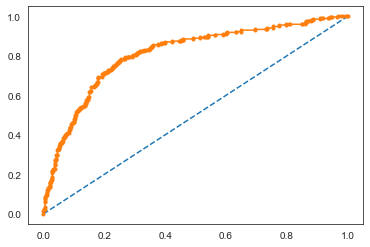
**macro avg 0.74 0.71 0.72 859**

**weighted avg 0.76 0.77 0.76 859**

#### ROC PLOT FOR TEST DATA:

**AUC: 0.811**

Figure ROC curve for test data RF

****

Train Data Accuracy: 0.80; Test Data Accuracy: 0.77;

Train Data Precision: 0.71; Test Data Precision: 0.67;

Train Data f1-score: 0.66; Test Data f1-score: 0.59;

Train Data Recall: 0.62; Test Data Recall: 0.53;

Train Data AUC: 0.852; Test Data AUC: 0.811;

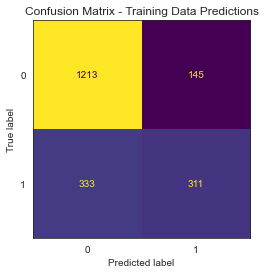
From observing the characteristics of the RF training & testing data set, RF model has better accuracy, precision, recall & f1 score than the CART model.

### ARTIFICIAL NEURAL NETWORK MODEL PERFORMANCE EVALUATION:

#### Confusion Matrix and Accuracy for the training data:

Accuracy for training data is **0.7612387612387612.**

Table 11 Confusion Matrix ANN for training data



#### Classification Report for the training data:

**precision recall f1-score support**

**0 0.78 0.89 0.84 1358**

**1 0.68 0.48 0.57 644**

**accuracy 0.76 2002**

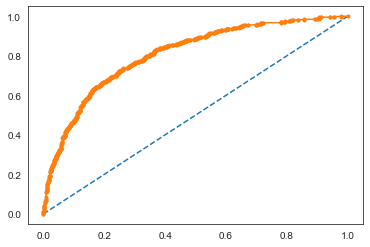
**macro avg 0.73 0.69 0.70 2002**

**weighted avg 0.75 0.76 0.75 2002**

#### ROC PLOT FOR TRAINING DATA:

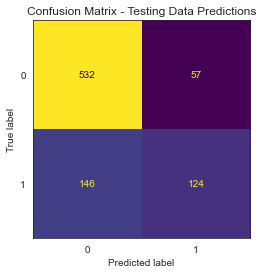
**AUC: 0.808**

Figure ANN ROC Curve for training data

****

#### Confusion Matrix and Accuracy for the testing data:

Table 12 Confusion Matrix For ANN Test data



Accuracy of test data is **0.7636786961583236.**

#### Classification Report for the testing data:

**precision recall f1-score support**

**0 0.78 0.90 0.84 589**

**1 0.69 0.46 0.55 270**

**accuracy 0.76 859**

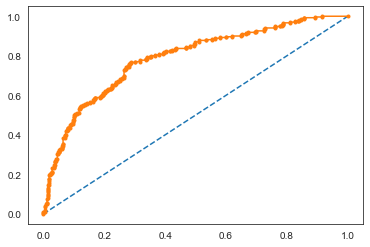
**macro avg 0.73 0.68 0.69 859**

**weighted avg 0.75 0.76 0.75 859**

#### ROC PLOT FOR TESTING DATA:

**AUC: 0.788**

Figure ROC curve ANN for Test data



Train Data Accuracy: 0.76; Test Data Accuracy: 0.76;

Train Data Precision: 0.68; Test Data Precision: 0.69;

Train Data Recall: 0.48; Test Data Recall: 0.46;

Train Data f1-score: 0.57; Test Data f1-score: 0.55;

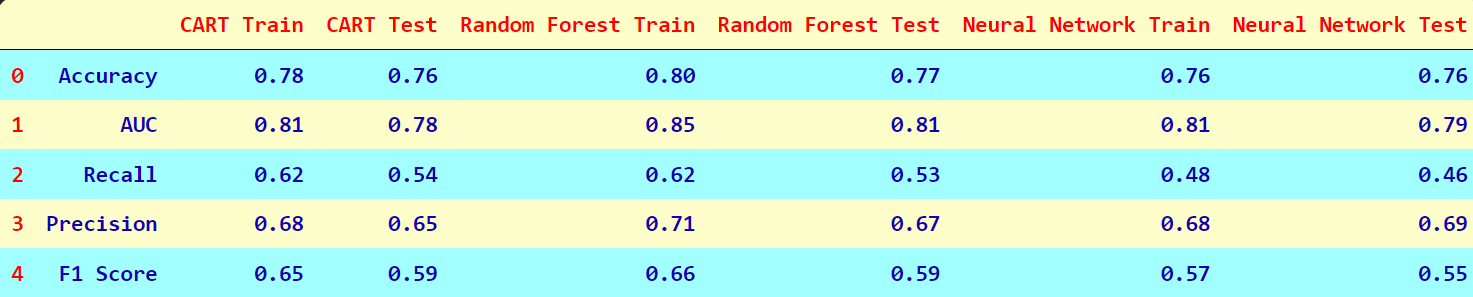
Train Data AUC: 0.808; Test Data AUC: 0.788;

From observing the characteristics of the ANN training & testing data set, RF model has better accuracy, precision, recall & f1 score than the model. So, we can finalise on RF Model.

## **2.4** Final Model: Compare all the models and write an inference which model is best/optimized.

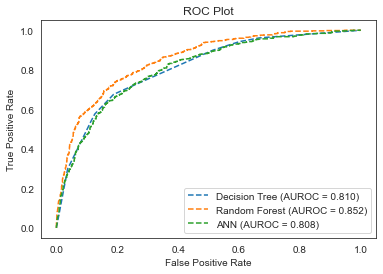
### Classification Report Comparison for all the three models:

Table Classification Report comparison



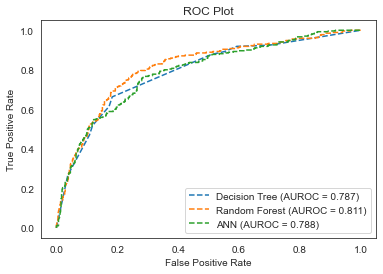
### ROC Curve for all the 3 models (Training data):

Figure ROC of training data for all three models



### ROC Curve for all the 3 models (Testing data):

Figure ROC curve for testing data of all Three Models



From observing the characteristics of the CART, RF, ANN training & testing data set, RF model has better accuracy, precision, recall & f1 score than the other two models. The Random Forest method has the best performance, i.e., best accuracy compared to all the three models. The percentage deviation between Training and Testing Dataset also is reasonably under control, classifying it as a good model among the three models.

## **2.5** Inference: Based on the whole Analysis, what are the business insights and recommendations.

1. The Basic objective for building the predictive model was to see & also classify if an insurance firm providing insurance is facing low/medium/high claim frequency.

2. The data had Outliers. After, performing & running the 3 models we inferred that the data is well balanced for conducting the models but, more data will assist & help understand to predict the models better. Past Data could also help in understanding & structuring the data to dive into deeper business problems. Overall, all the models are stable enough for making any future predictions because there is no overfitting.

3. Claims are higher for ASIA destination; the management could take actions & follow stipulated protocols before structuring a policy in ASIA. This, could be done by increasing rates via Premium on the policy for recovering the Cost which has been claimed.

4. The Management should make more Customers opt for an insurance policy more affordable & increase complexion of the policy in such a way that more customers would choose the insurance policy for the competitive rate. This would attract more customers & also certainly reduce fraud claims by the set of complexions.

5. As we could observe from the data maximum of the insurance claim has been done by online channel than offline channel, reason being more convenient & good experience which is a good factor generating profits. But, subsequently, management should also promote offline channels by giving additional promotions, offers, low cost than online to pull customers which could generate more customer base.

6. We could observe that most of the Sales was generated from Type Sales than Airlines & also we could observe that Insurance Claim are more at Airlines than Agencies. This could be because of more customer satisfaction & customer trust at Airlines than at Agencies. This could be put into better use by suggestive selling insurance plans at Airlines. Management could deploy better customer service at Agencies to match the likeliness of Airlines which could give better Sales Performance. Low performing Agencies such as JZI, CWT could be given more promotional drives for better sales conversion rates.

7. Customized Plans have been chosen more compared to other plans. Targets & incentives could be given to promote Silver Plans & Gold plans which would generate more sales via Promo offers, Ad Campaigns, Marketing drives etc.

8. Management should focus more on Customer Satisfaction, Customer Claim Turn Around Times, methods to actively reduce false transactions via strong Artificial Intelligence Methods. Focusing on making the process efficient & affordable & reducing overall operational costs.