**DATA MINING PROJECT**



**CLUSTERING | CART | RF | ANN**

**Gaurav Sood**

**PGP-DSBA Online July 2021**

**Date: 21st November, 2021**

**Table of Contents**

**Problem 1: Clustering…………………………………………………………………………….6**

**1.1** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)………………………………………7

**1.2**  Do you think scaling is necessary for clustering in this case? Justify…………13

**1.3** Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them…………………………………14

**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………………………………………………………16

**1.5** Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters……………………………………………..17

**Problem 2: CART-RF-ANN………………………………………………………………………19**

**2.1** Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)……………………………………………………………...20

**2.2** Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network………………………………………………………………...…30

**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………………………………………………...……32

**2.4** Final Model: Compare all the models and write an inference which model is best/optimized…………………………………………………………………………………….....37

**2.5** Inference: Based on the whole Analysis, what are the business insights and recommendations…………………………………………………………………………………...37

**List of Tables**

Table no.1: Dataset Sample………………………………………………………………………………………………………………6

Table no.2: Data type and non-null value…………………………………………………………………………………………..6

Table no.3: Checking for null values…………………………………………………………………………………………………..6

Table no.4: Data Description…………………………………………………………………………………………………………….7

Table no.5: Scaled Data Sample……………………………………………………………………………………………………...14

Table no.6: Hierarchical Clustering Sample Dataset…………………………………………………………………………15

Table no.7: Grouping by Clusters (mean)………………………………………………………………………………………..15

Table no.8: K-Means Clustering Sample Dataset………………………………………………………………………………17

Table no.9: Grouping by K-Means clusters……………………………………………………………………………………….17

Table no.10: Dataset Sample…………………………………………………………………………………………………………...19

Table no. 11: Data type and non-null value………………………………………………………………………………………19

Table no. 12: Data Description…………………………………………………………………………………………………………20

**List of Figures**

Figure no. 1: ………………………………………………………………………………………………………………………………………8

Figure no. 2: ………………………………………………………………………………………………………………………………………8

Figure no. 3: ………………………………………………………………………………………………………………………………………8

Figure no. 4: ………………………………………………………………………………………………………………………………………9

Figure no. 5: ………………………………………………………………………………………………………………………………………9

Figure no. 6: ………………………………………………………………………………………………………………………………………9

Figure no. 7: ……………………………………………………………………………………………………………………………………10

Figure no. 8: ……………………………………………………………………………………………………………………………………10

Figure no. 9: ……………………………………………………………………………………………………………………………………12

Figure no. 10: ………………………………………………………………………………………………………………………………….13

Figure no. 11: ………………………………………………………………………………………………………………………………….14

Figure no. 12: ………………………………………………………………………………………………………………………………….15

Figure no. 13: ………………………………………………………………………………………………………………………………….16

Figure no. 14: ………………………………………………………………………………………………………………………………….17

Figure no. 15: ………………………………………………………………………………………………………………………………….21

Figure no. 16: ………………………………………………………………………………………………………………………………….22

Figure no. 17: ………………………………………………………………………………………………………………………………….23

Figure no. 18: ………………………………………………………………………………………………………………………………….23

Figure no. 19: ………………………………………………………………………………………………………………………………….24

Figure no. 20: ………………………………………………………………………………………………………………………………….25

Figure no. 21: ………………………………………………………………………………………………………………………………….26

Figure no. 22: ………………………………………………………………………………………………………………………………….27

Figure no. 23: ………………………………………………………………………………………………………………………………….28

Figure no. 24: ………………………………………………………………………………………………………………………………….29

Figure no. 25: ………………………………………………………………………………………………………………………………….30

Figure no. 26: …………………………………………………………………………………………………………………………………33

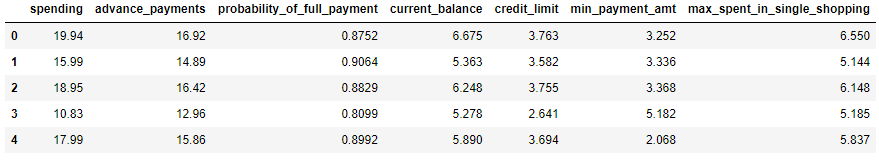
Figure no. 27: ………………………………………………………………………………………………………………………………….34

Figure no. 28: ………………………………………………………………………………………………………………………………….36

**Problem 1: Clustering**

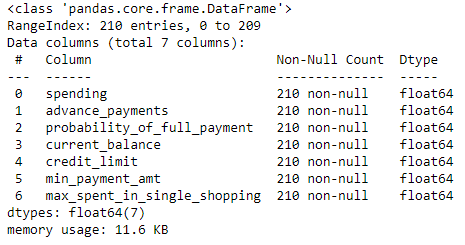
A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

**Sample of the Dataset**



**Table no. 1: Dataset Sample**

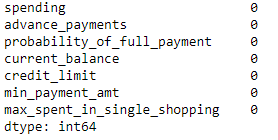
**Exploratory Data Analysis**



**Table no.2: Data type and non-null value**

The dataset contains 210 rows across 7 columns, with all entries being of float type.

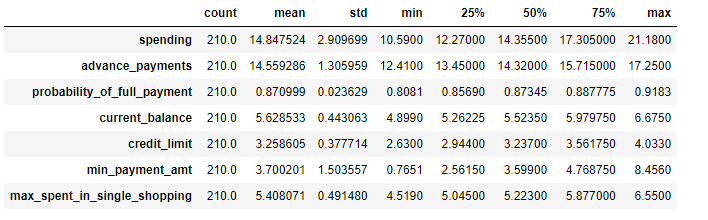
**Checking for Null Values**

****

**Table no.3: Checking for null values**

There are no null values in the given dataset.

**Data Description**



**Table no. 4: Data Description**

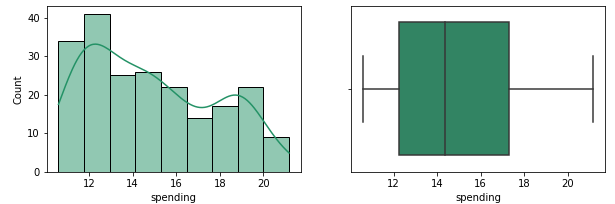
The data dictionary for Market Segmentation is given as below:

1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current\_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)

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

We start off with Univariate analysis for each column in the dataset.

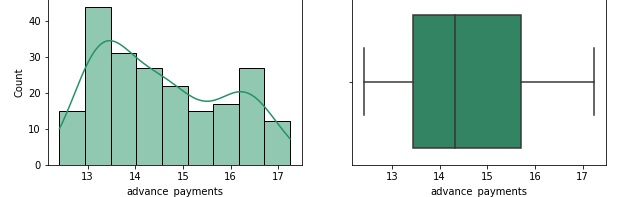
1. **Univariate Analysis**
2. **Spending**

****

**Figure no.1: Histogram and Boxplot for Spending**

The data points are slightly left skewed and have no outliers in the column.

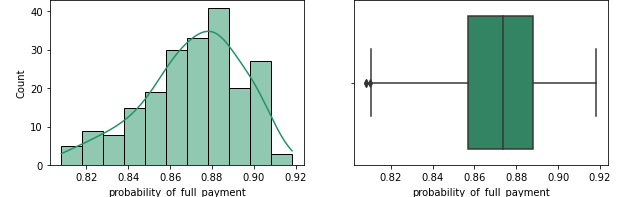
1. **Advance Payments**

****

**Figure no.2: Histogram and Boxplot for Advance Payments**

The data points are slightly left skewed and have no outliers in the column.

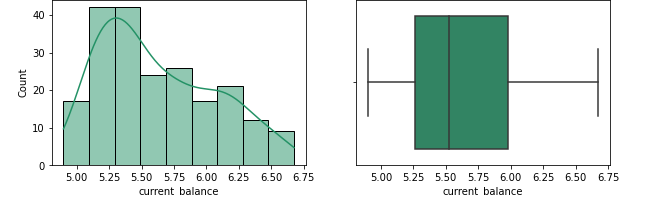
1. **Probability of Full Payment**

****

**Figure no.3: Histogram and Boxplot for Probability of Full Payment**

The data points are right skewed and there is a presence of outliers in the column.

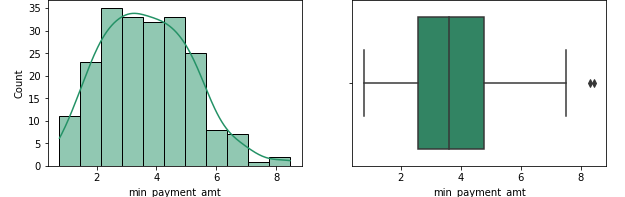
1. **Current Balance**

****

**Figure no.4: Histogram and Boxplot for Current Balance**

The data points are left skewed and have no outliers in the column.

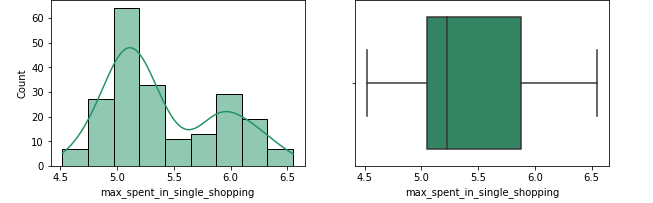
1. **Minimum Payment Amount**

****

**Figure no.5: Histogram and Boxplot for Minimum Payment Amount**

The data points are left skewed and there are outliers present in the column.

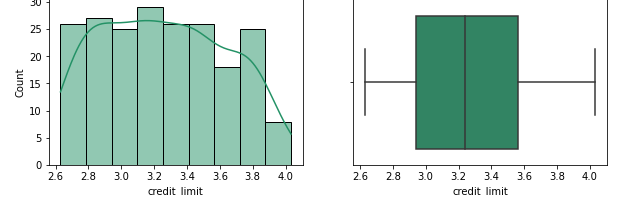
1. **Max Spend in Single Shopping**

****

**Figure no.6: Histogram and Boxplot for Max Spent in Single Shopping**

The data points are normally distributed and have no outliers in the column.

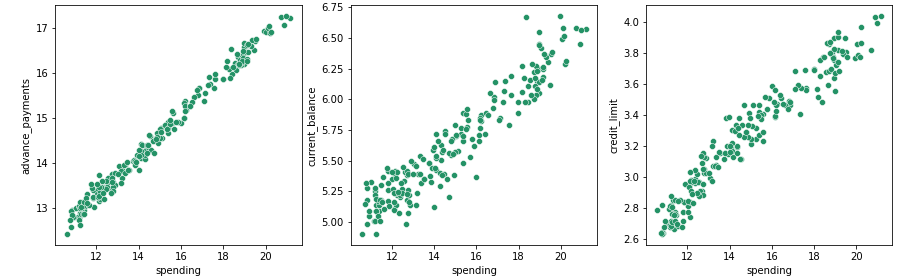
1. **Credit Limit**

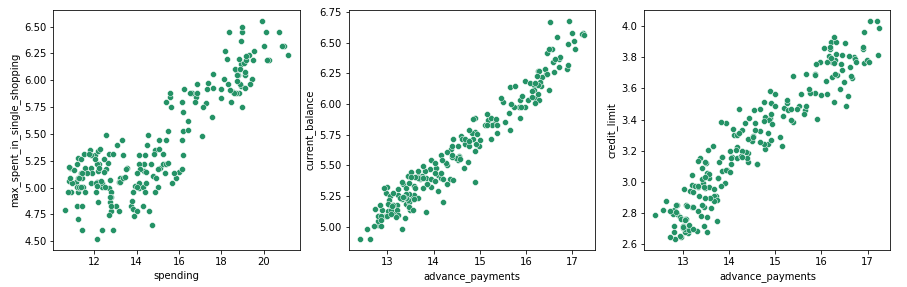
****

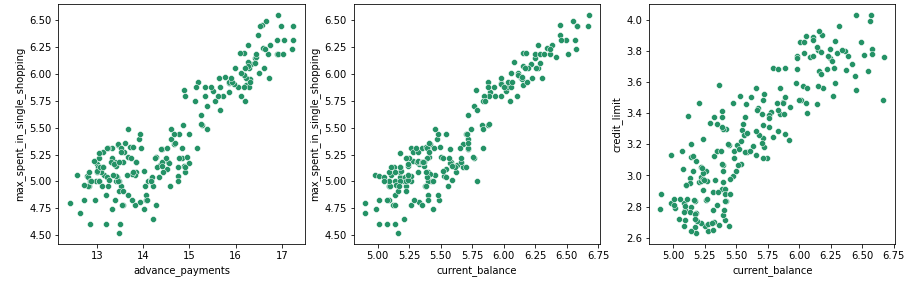
**Figure no.7: Histogram and Boxplot for Credit Limit**

The data points are normally distributed and have no outliers in the column.

1. **Bivariate Analysis**

****

****

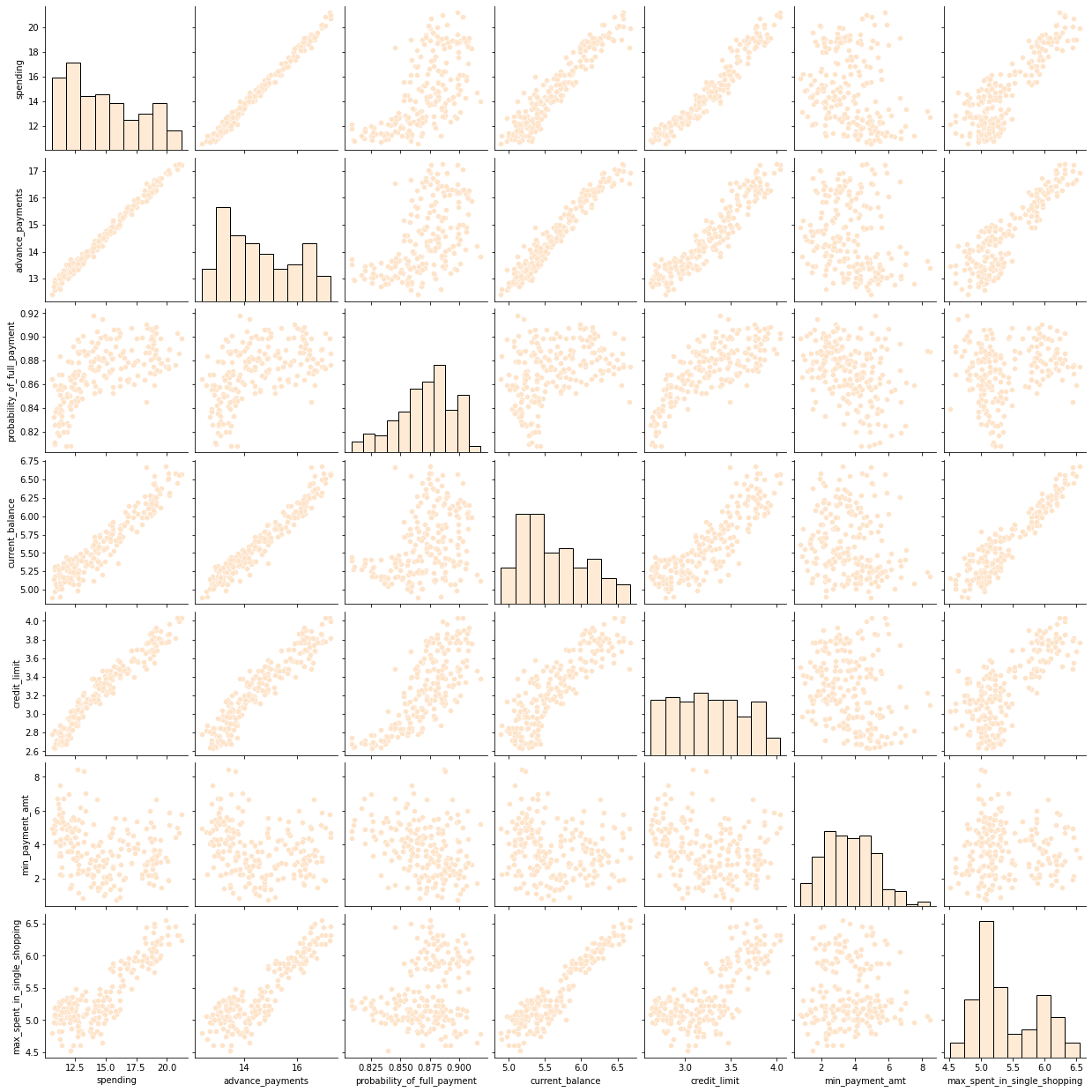
****

**Figure no.8: Scatterplot of Variables**

The above plot indicates that:

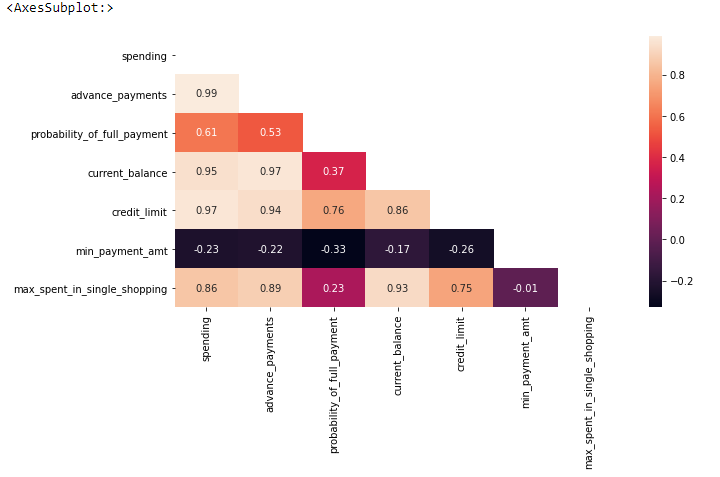
* Spending increases as credit balance and credit limit increases.
* As current balance and credit limit increase, the customer starts with advance payment.
* Customers with higher current balance go with maximum spending in a single shopping.

1. **Multivariate Analysis**



**Figure no.9: Pairplot of all Variables**

**Correlation Plot**



**Figure no.10: Correlation Plot of the Variables**

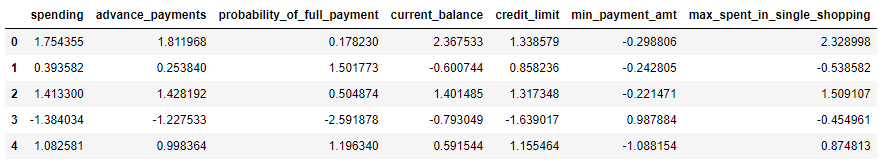
From the above correlation plot, we can say that:

* Spending is highly correlated with advance payment, current balance, credit limit and max spend in single shopping.
* Advance payment is highly correlated with current balance, credit limit and max spending in single shopping.
* Current balance is highly correlated with credit limit and max spend in single shopping.
* Credit limit is highly correlated with max spend in single shopping.

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

Yes, scaling is necessary for clustering. For implementation and better performance of the machine learning model, the data should be scaled in single unit form.

For this business problem we proceed with the Standard scalar technique.



**Table no. 5: Scaled Data Sample**

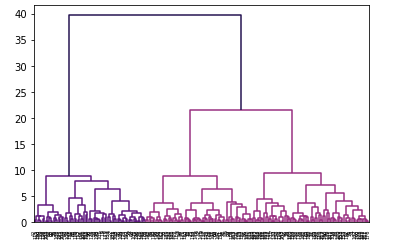
The dataset is now scaled with mean value = 0 and standard deviation = 1.

We will make use of this scaled data for cluster formation and model building.

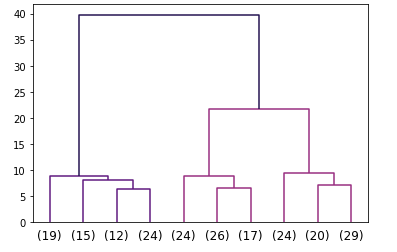
**1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them**

In hierarchical clustering, different linkage methods are used to find the optimum number of clusters. For our problem dataset we go with the Ward linkage method.

After applying Ward linkage method, we get a dendrogram as below:



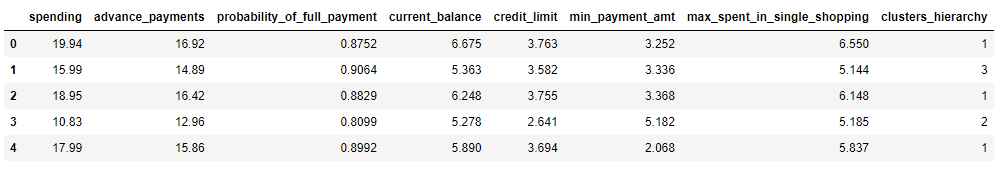
**Figure no.11: Dendrogram for Hierarchical Clustering**

****

**Figure no.12: Dendrogram for Optimum Clustering**

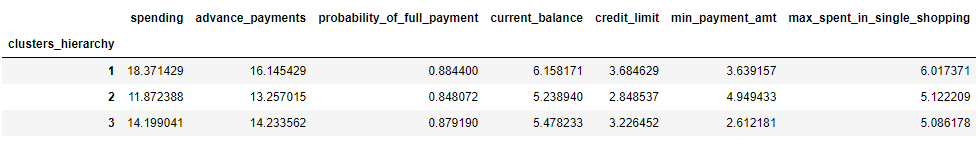
From the above figure, we will go for 3 clusters based on the Euclidean distance between clusters.

**Sample Dataset after applying Hierarchical Clustering**



**Table no. 6: Hierarchical Clustering Sample Dataset**

**Grouping by Clusters**



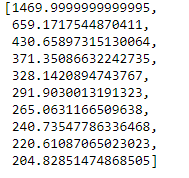
**Table no. 7: Grouping by Clusters (Mean)**

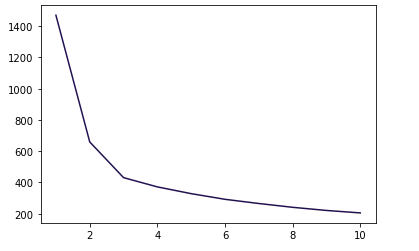
After grouping by clusters and applying mean function, the customers are divided into 3 categories, namely:

1. High Spender: Those having high spending capabilities.
2. Medium Spender: Those having moderate spending capabilities.
3. Low Spender: Those having low spending capabilities.

**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 clustering, in order to find the optimum number of clusters we use the elbow curve and silhouette score. After applying K-Means clustering we get a wss plot as below:





**Figure no.13: Elbow Curve**

From fig no. 13, we can see that after 430.66 the values are reducing in small range.

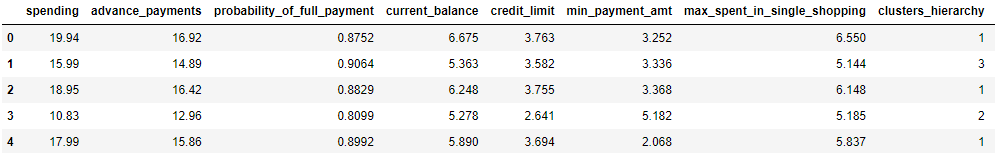
Similarly in the elbow plot illustrated in fig no. 13, after the 3rd data point the curve looks quite smooth. So we can safely say that the optimum number of clusters should be 3.

**Silhouette Score**

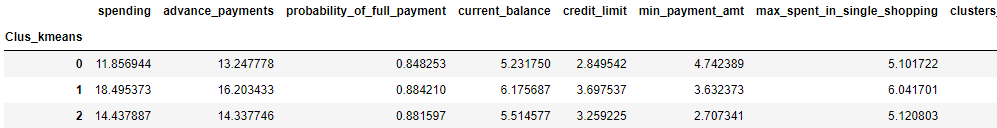


0.40 is the silhouette score for 3 clusters which is good enough to finalize optimum clusters.

**Sample of Dataset after K-Means Clustering**



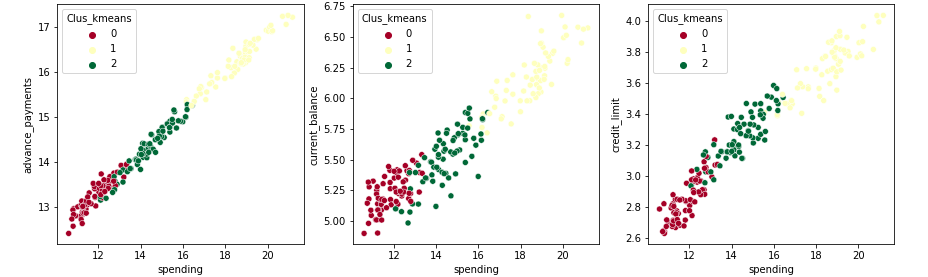
**Table no. 8: Sample data after K-Means Clustering**

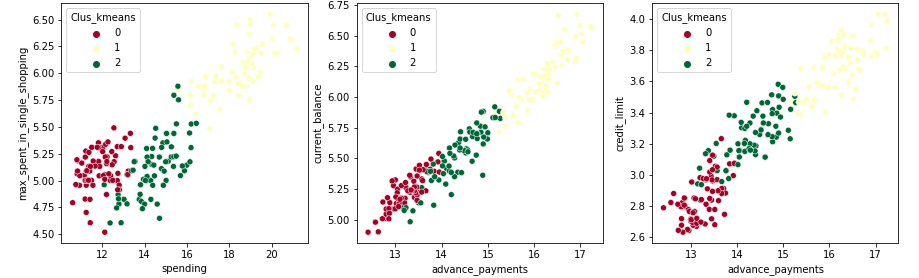
****

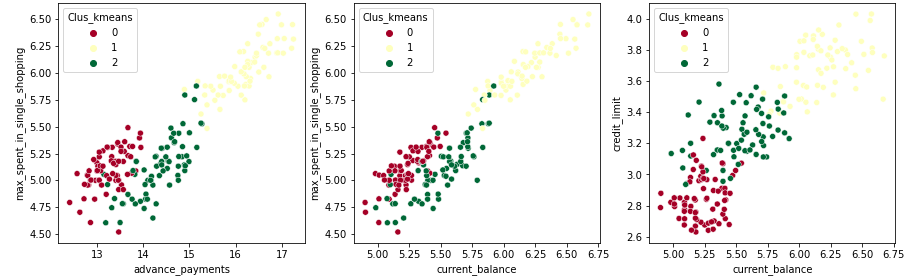
**Table no. 9: Grouping by K-Means Clusters**

Grouping with K-Means clustering has given us some comparable results to Hierarchical clustering.

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





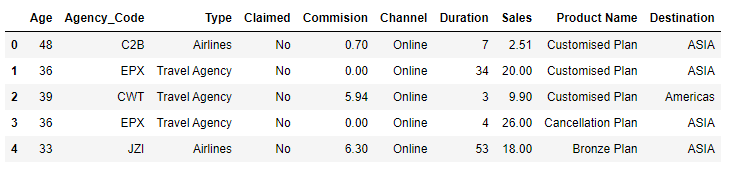
****

**Figure no.14: Scatterplot**

**Problem 2: CART-RF-ANN**

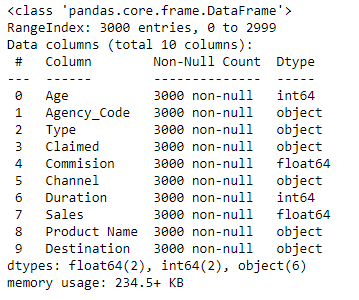
An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models'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

**Sample of the Dataset:**



**Table no. 10: Dataset Sample**

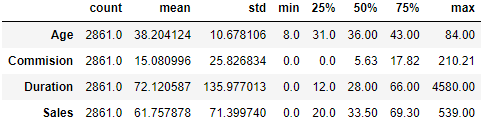
**Exploratory Data Analysis**



**Table no. 11: Data type and non-null value**

The dataset contains 3000 rows across 10 columns, with six of them being of object type and two each of float and integer typing.

**Data Description**

****

**Table no. 12: Data Description**

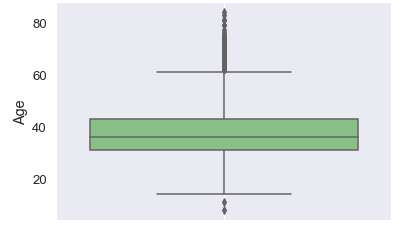
Attribute Information:

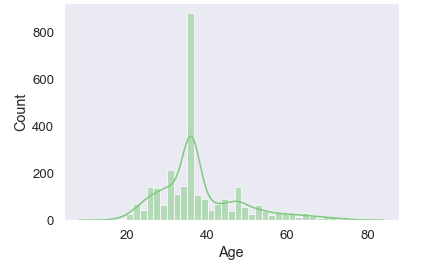
1. Target: Claim Status (Claimed)  
2. Code of tour firm (Agency\_Code)  
3. Type of tour insurance firms (Type)  
4. Distribution channel of tour insurance agencies (Channel)  
5. Name of the tour insurance products (Product)  
6. Duration of the tour (Duration in days)  
7. Destination of the tour (Destination)  
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)  
9. The commission received for tour insurance firm (Commission is in percentage of sales)  
10.Age of insured (Age)

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

We first proceed with univariate analysis for each of the columns in the dataset.

1. **Univariate Analysis**
2. **Age**

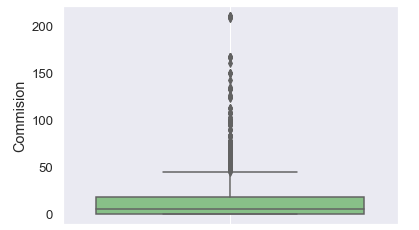
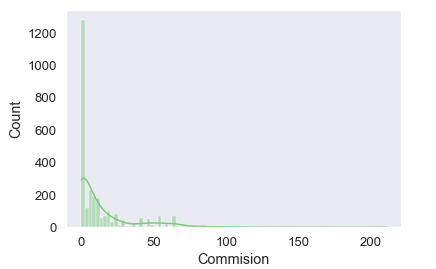
****

****

**Figure no.15: Histogram and Boxplot for Age**

The data points are normally distributed and there are outliers in the column.

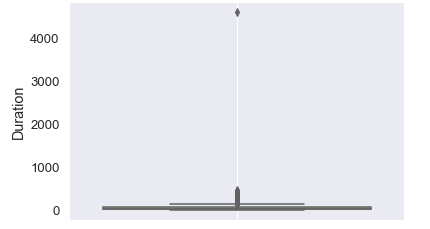
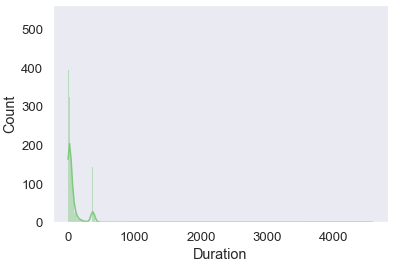
1. **Commission**

**** ****

**Figure no.16: Histogram and Boxplot for Commision**

The data points are left skewed and containing outliers in the column.

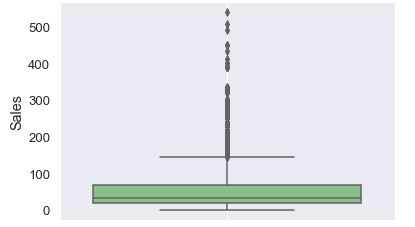
1. **Duration**

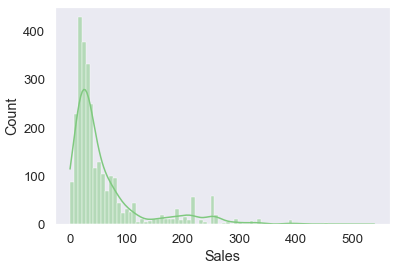
**** ****

**Figure no.17: Histogram and Boxplot for Duration**

The data points are left skewed and having outliers in the column.

1. **Sales**

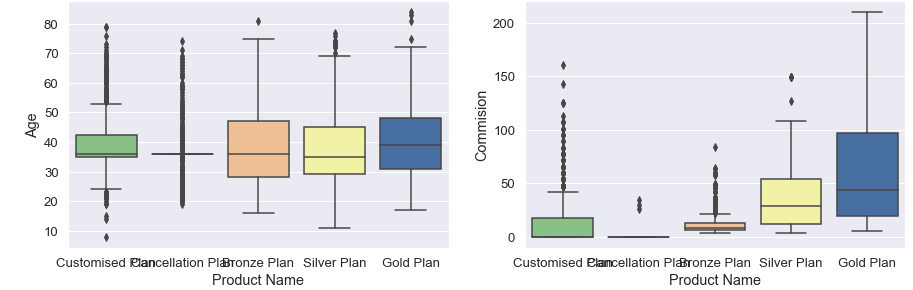
****

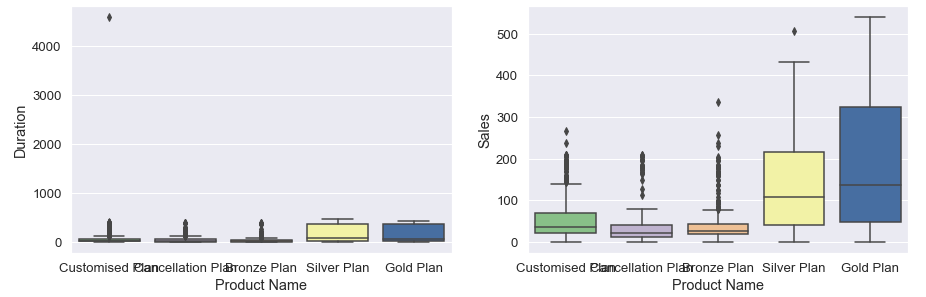
****

**Figure no.18: Histogram and Boxplot for Sales**

The data points are left skewed and there is a presence of outliers in the column.

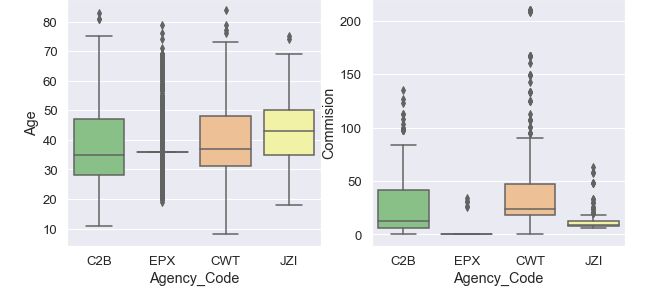
1. **Bivariate Analysis**

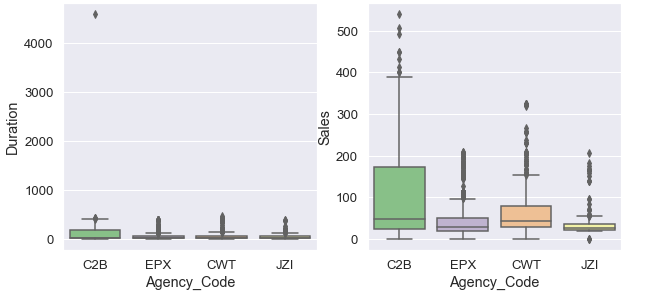
****

****

**Figure no.19: Boxplot of Product Name vs Continuous Columns**

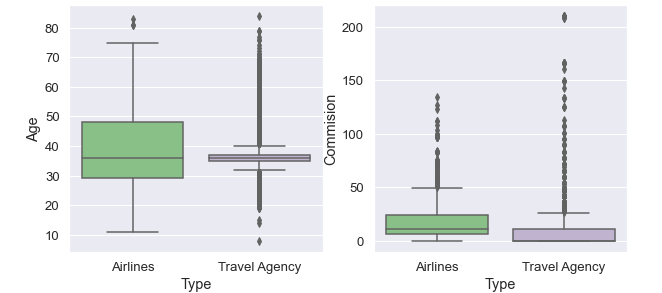
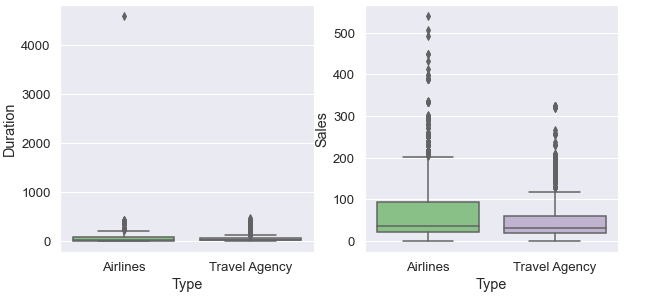
Sales and subsequently Commission are quite large for Gold Plan as compared to other products.

****

****

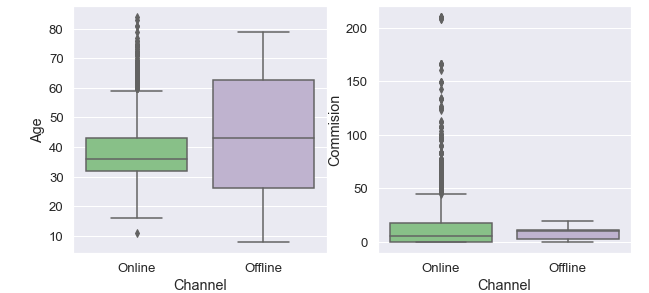
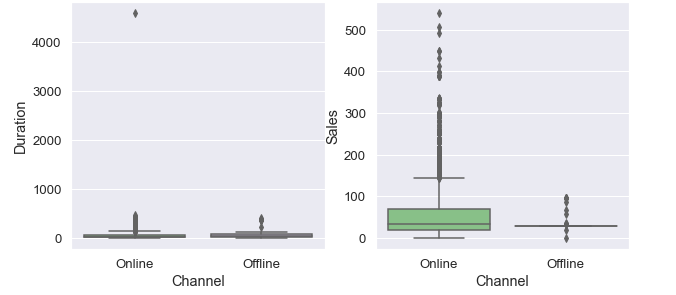
**Figure no.20: Boxplot of Agency\_Code vs Continuous Columns**

C2B and CWT Agency Code get more Commission as compared to the other two.

**** ****

**Figure no.21: Boxplot of Type vs Continuous Columns**

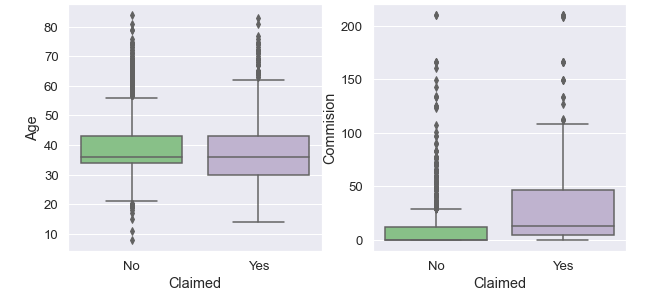
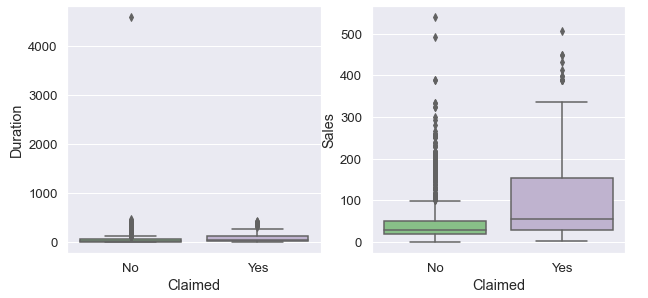
Ages 30 to 50 prefer Airlines mode over Travel Agency.

**** ****

**Figure no.22: Boxplot of Channel vs Continuous Columns**

Sales and Commission are higher for Online channels as compared to Offline channels.

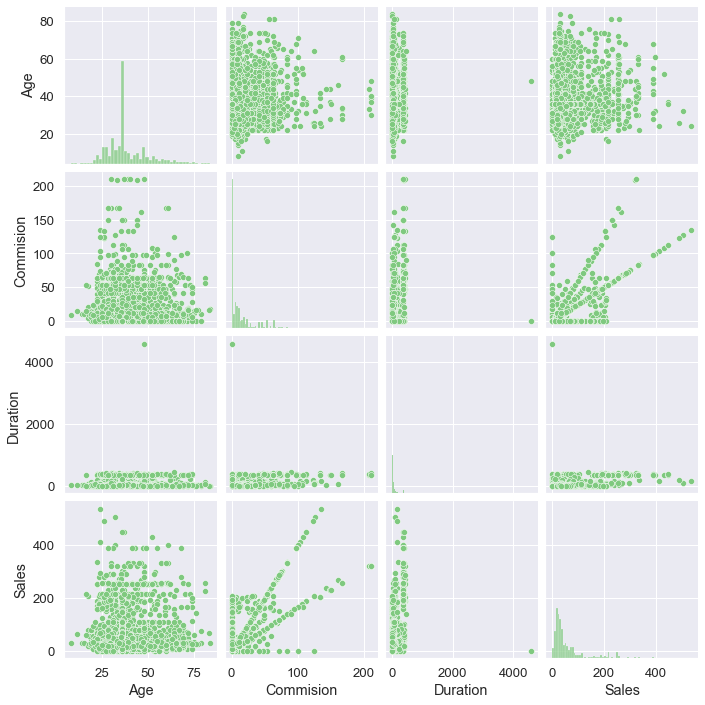
Customers mostly use Offline channels.

**** ****

**Figure no.23: Boxplot of Claimed vs Continuous Columns**

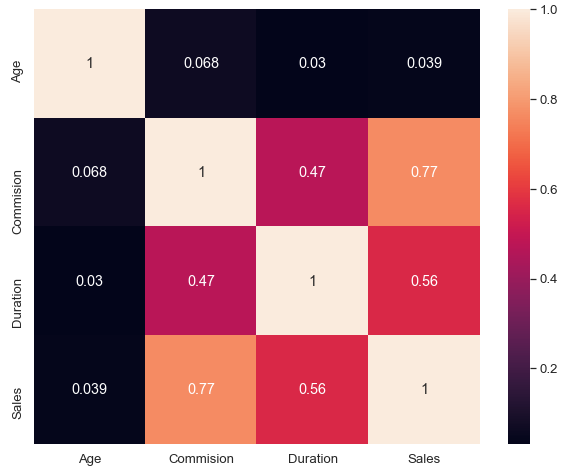
Those who have claimed ‘Yes’ have higher sales and subsequently, commission than those who claimed ‘No’.

1. **Multivariate Analysis**

****

**Figure no.24: Pairplot of the Variables**

**Correlation Plot**



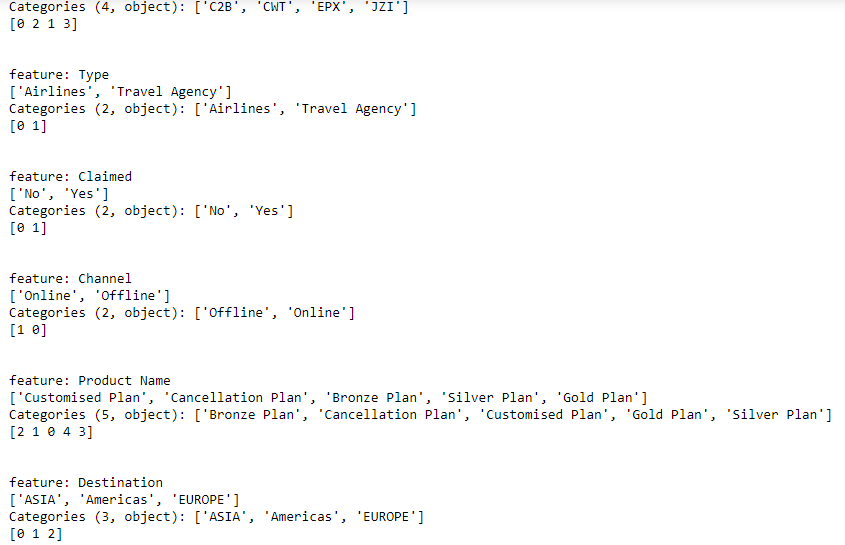
**Figure no.25: Correlation Plot of the Variables**

From the above correlation plot, we can infer that:

* Continuous columns are showing less correlation with each other.
* Sales and commission are highly correlated.
* As sales increase, commission also increases.

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

Converting the categorical data:



Then we create two variables x and y, where x is a dependent variable and y is an independent variable.

Let y be the claimed column and x be the rest of the columns after dropping claimed column.

Splitting the data into train and test sample:



As we can see, the data has been split into x and y; with

**Test size = 0.3**

**Random state = 1**

**CART Model**

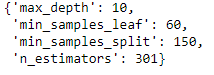
Training and testing our model on decision tree model, we get:

****

After applying grid search and cross validation, we have got the best parameter for our CART model.

**Random Forest Model**

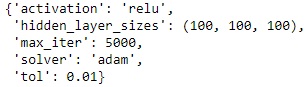
Training and testing our model on random forest model, we get:

****

This is the best parameter for random forest model.

**Artificial Neural Network**

Training and testing our model on artificial neural network model, we get:

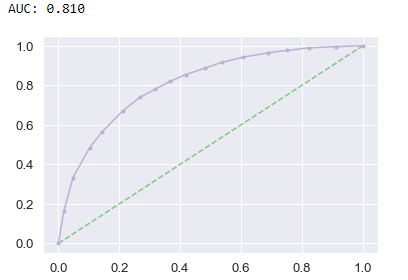
****

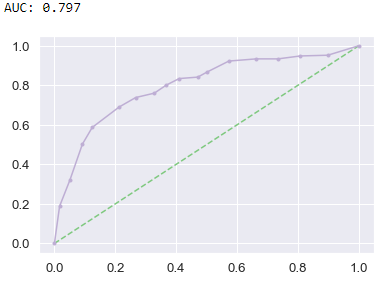
This is the best parameter for artificial neural network.

**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**

Checking all conclusive parameters:

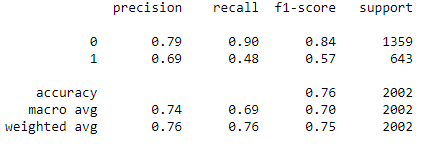




**Figure no.26: ROC\_AUC\_Curve for CART Model**

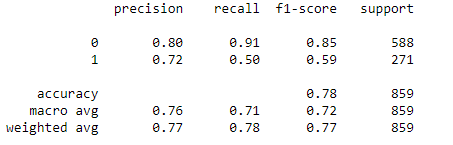
From the above figure, we can see that the AUC score for both train and test model are quite close to each other.

**Classification Report and Confusion Matrix for Train data sample:**



The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the train data sample.

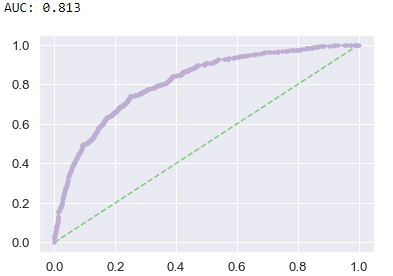
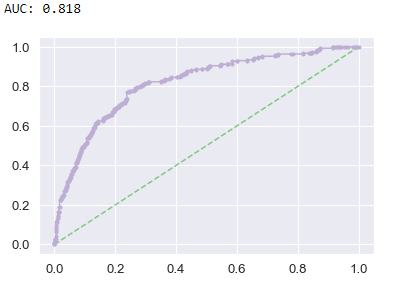
**Classification Report and Confusion Matrix for Test data sample:**



The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the test data sample.

**Random Forest Model**

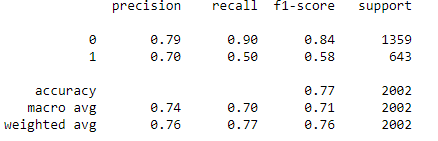
Checking all conclusive parameters:

**** ****

**Figure no.27: ROC\_AUC\_Curve for Random Forest Model**

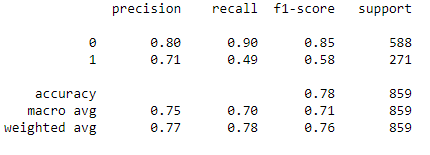
From the above figure, we can see that the AUC score for both train and test model are very close to each other.

**Classification Report and Confusion Matrix for Train data sample:**



The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the train data sample.

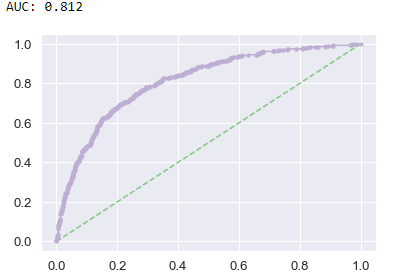
**Classification Report and Confusion Matrix for Test data sample:**

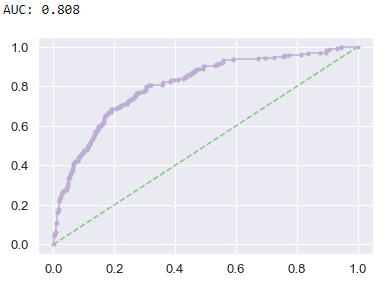


The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the test data sample.

**Artificial Neural Network Model**

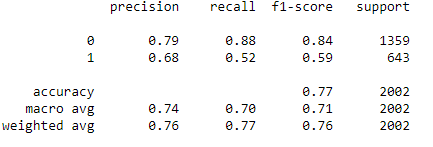
Checking all conclusive parameters:





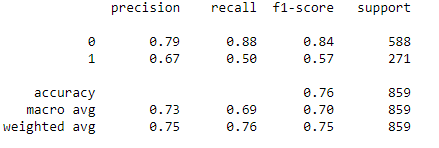
**Figure no.28: ROC\_AUC\_Curve for Artificial Neural Network Model**

**Classification Report and Confusion Matrix for Train data sample:**



The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the train data sample.

**Classification Report and Confusion Matrix for Test data sample:**



The model is performing well for Claimed ‘no’ but showing poor results for claimed ‘yes’ in the test data sample.

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

From the above models and testing, we can see that all the models are showing quite similar results in all aspects.

However, the Artificial Neural Network model is a bit more accurate for our problem as compared to CART and Random Forest models.

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

* Agency\_Code is the most relevant column in the dataset to fetch insight.
* C2B Agency Code impacts Sales massively. More analysis is needed in order to implement the same pattern to other agencies.
* Sales and Commission are highly correlated. To increase Sales one must focus on Commission given to different agency.
* Age and Duration column do not affect Claimed as much, so we can disregard these columns.
* There is a need for more relevant data for further model building and improving machine performance.