DATA MINING PROJECT

Shankar S

Report on Bank Marketing Data and Insurance Data which is performed by Clustering and Classification Techniques. Also techniques like Decision trees, Random Forest and Artificial Neural Network is used to compare which model works more effectively with the Dataset.

PGP – DSBAONLINE

BATCH: APRIL 2021

DATE: 29/08/2021

Table of Contents

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.

PROBLEM 2: CART-RF-ANN

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, Bivariate, and multivariate analysis).

The sample of the data is displayed below.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

The description of the data is

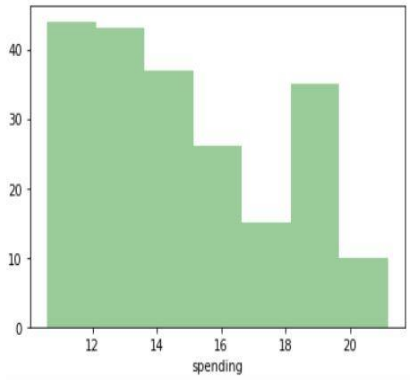
	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

The info about the given Dataset is displayed below.

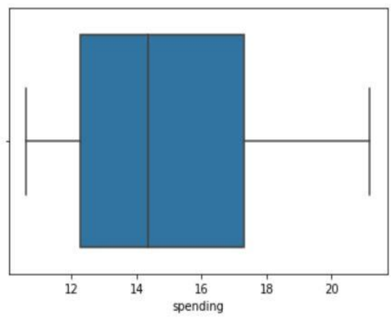
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

Univariate Analysis has been performed for all the features and the results is displayed below.

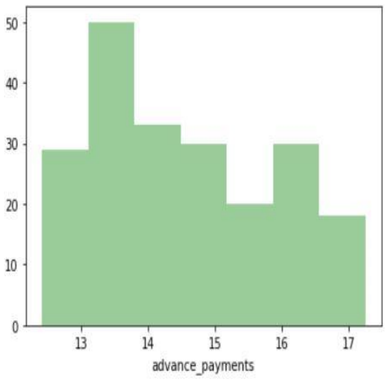
```
Description of spending
----------------------------
count    210.000000
mean      14.847524
std        2.909699
min       10.590000
25%       12.270000
50%       14.355000
75%       17.305000
max       21.180000
Name: spending, dtype: float64
----------------------------
```



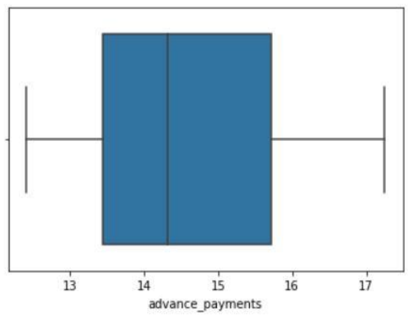
BoxPlot of spending



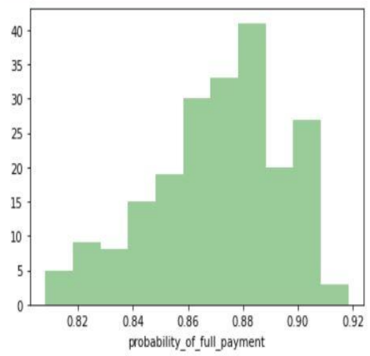

```
Description of advance_payments
----------------------------------
count    210.000000
mean      14.559286
std        1.305959
min       12.410000
25%       13.450000
50%       14.320000
75%       15.715000
max       17.250000
Name: advance_payments, dtype: float64
----------------------------------
```



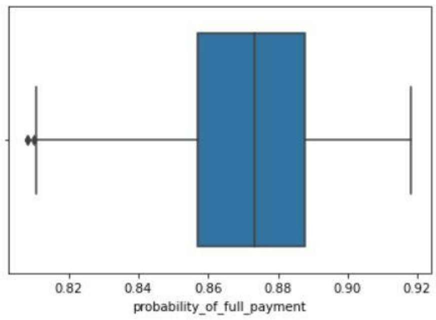
BoxPlot of advance_payments



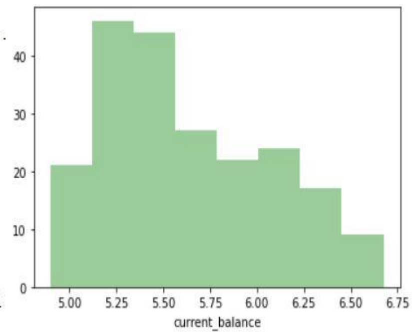

```
Description of probability_of_full_payment
--------------------------------------------
count    210.000000
mean       0.870999
std        0.023629
min        0.808100
25%        0.856900
50%        0.873450
75%        0.887775
max        0.918300
Name: probability_of_full_payment, dtype: float64
--------------------------------------------
```



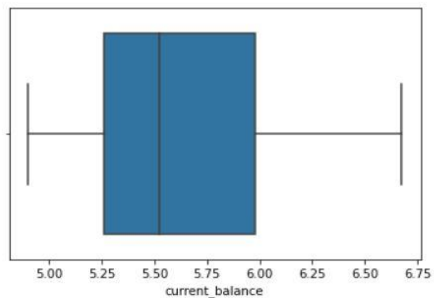
BoxPlot of probability_of_full_payment



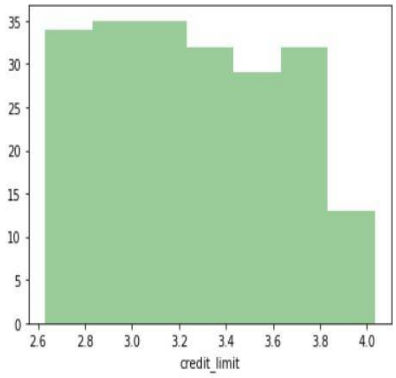

```
Description of current_balance
----------------------------------
count    210.000000
mean       5.628533
std        0.443063
min        4.899000
25%        5.262250
50%        5.523500
75%        5.979750
max        6.675000
Name: current_balance, dtype: float64
----------------------------------
```



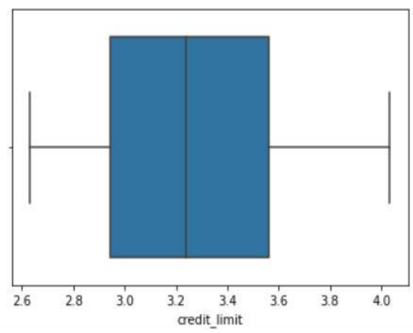
BoxPlot of current_balance



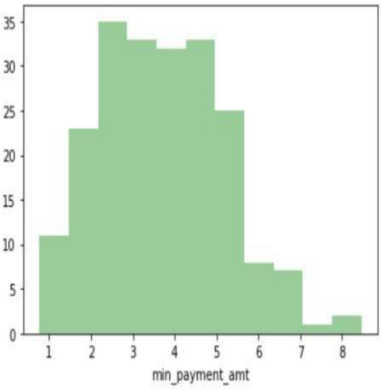
```
Description of credit_limit
----------------------------------
count    210.000000
mean       3.258605
std        0.377714
min        2.630000
25%        2.944000
50%        3.237000
75%        3.561750
max        4.033000
Name: credit_limit, dtype: float64
----------------------------------
```



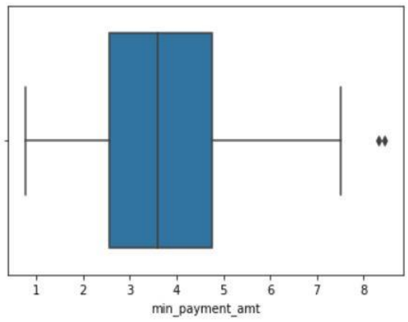
BoxPlot of credit_limit



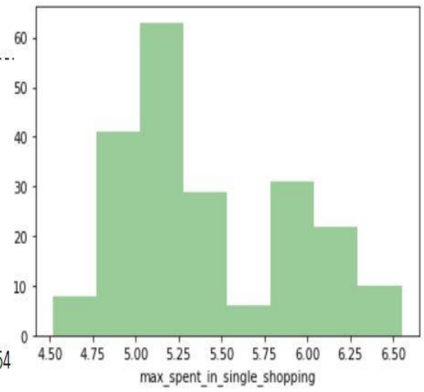

```
Description of min_payment_amt
--------------------------------------
count    210.000000
mean       3.700201
std        1.503557
min        0.765100
25%        2.561500
50%        3.599000
75%        4.768750
max        8.456000
Name: min_payment_amt, dtype: float64
--------------------------------------
```



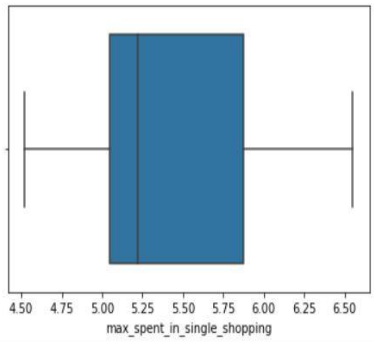
BoxPlot of min_payment_amt




```
Description of max_spent_in_single_shopping
---------------------------------------------------
count    210.000000
mean       5.408071
std        0.491480
min        4.519000
25%        5.045000
50%        5.223000
75%        5.877000
max        6.550000
Name: max_spent_in_single_shopping, dtype: float64
---------------------------------------------------
```



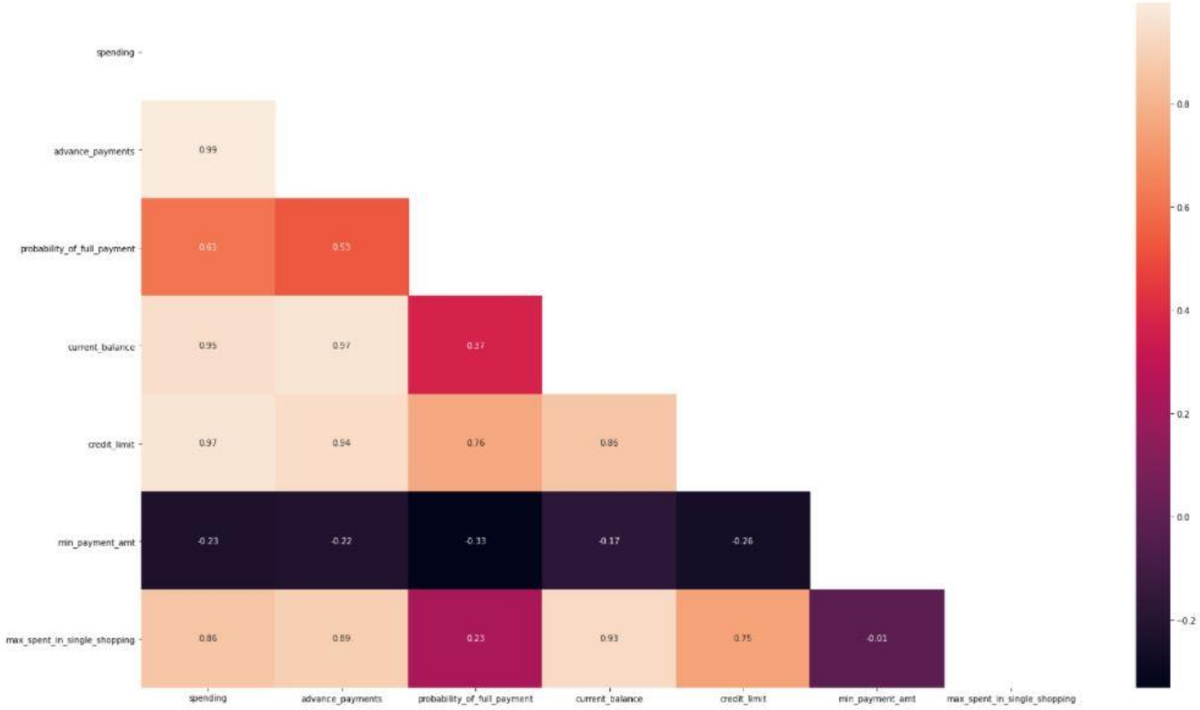
BoxPlot of max_spent_in_single_shopping



By viewing this we can analyse the following.

- The minimum amount of spending spent by a customer is 10590 and the maximum amount of spending spent is 21180.
- The minimum amount of advance payments done by a customer is 1241 and the maximum amount of advance payment paid by a customer is 1725 with an average of 1432.
- We can find outliers in probability_of_full_payment and min_payment_amt.

Multivariate Analysis has been performed for the given Dataset and the result is shown below.



We can see that the following features have strong correlation with other features.

- Advance payments is highly correlated with spending.
- Max spent in single shopping is highly correlated with spending, advance payments, current balance and credit limit.
- Credit limit is highly correlated with spending, advance payments, current balance.

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

We can observe from the dataset description that the range of all the features are in a different scale. Clustering is very sensitive to outliers. In this case Scaling is necessary for the given Dataset so that optimum clusters can be defined.

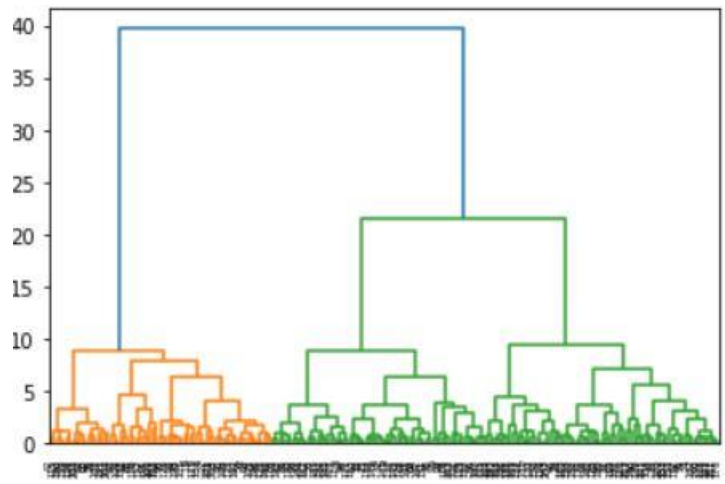
	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.5900	12.27000	14.35500	17.305000	21.1800
advance_payments	210.0	14.559286	1.305959	12.4100	13.45000	14.32000	15.715000	17.2500
probability_of_full_payment	210.0	0.870999	0.023629	0.8081	0.85690	0.87345	0.887775	0.9183
current_balance	210.0	5.628533	0.443063	4.8990	5.26225	5.52350	5.979750	6.6750
credit_limit	210.0	3.258605	0.377714	2.6300	2.94400	3.23700	3.561750	4.0330
min_payment_amt	210.0	3.700201	1.503557	0.7651	2.56150	3.59900	4.768750	8.4560
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.5190	5.04500	5.22300	5.877000	6.5500

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

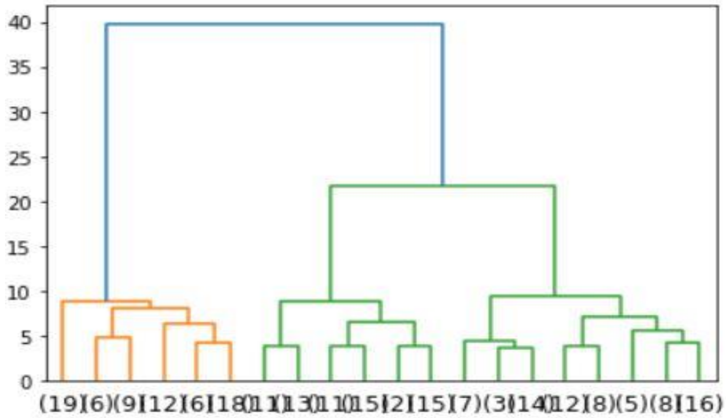
Scaling of the data is performed and Heirarchial Clustering is implemented for the scaled Data.

In this case we are using **'Ward Method'** to calculate the distance between the clusters. Ward's linkage is Similar to group average and centroid distance. It joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from the general approach.

A dendrogram is a treelike diagram that summarizes the process of clustering. Dendogram has been formed for the scaled data after performing Heirarchial Clustering and the output has been shown below.



The dendogram for the last 20 is shown below.



The optimum number of clusters formed after performing Heirarchial Clustering is ***Three*** which can be indentified from the Dendogram.

By using Dendogram we can also analyse the distance between the records which has been formed by performing Heirarchial Clustering.

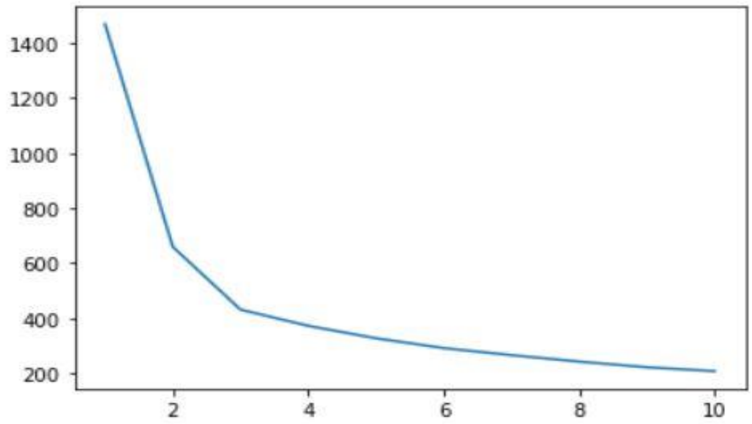
1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-Means is a non-hierarchical approach to forming good clusters is to pre-specify a desired number of clusters, k. The '**means**' in the K-means refers to averaging of the data; that is, finding the centroid.

K-Means is performed on the scaled Data and the inertia is calculated for the desired number of clusters.

```
[1469.9999999999995,
 659.1717544870411,
 430.65897315130064,
 371.6531439995162,
 326.3228713996129,
 290.628393695754,
 264.8862088334804,
 241.44962458453278,
 220.87269563766083,
 206.74286678894833]
```

Once the inertia has been calculated for number of clusters from 1 to 10 elbow curve is drawn to calculate the optimum value of clusters and is shown below.



INFERENCE:

The optimum number of clusters is identified by analysing the Inertia and elbow curve.

The ideal number of clusters is **Three** which is analysed from the Elbow curve.

As we can see there is a significant amount of drop when the clusters is changed from 1, 2 and 3. But there is no significant amount drop when the clusters is changed from 3 to 4. Also the Inertia for the number of clusters = 1 is 1469.99 and the inertia for the number of clusters = 2 is 659.1717 and the inertia for the number of clusters = 3 is 430.65. But the inertia of number of clusters = 4 is 371.65. We can able to see that there is a significant amount of drop of Inertia from number of clusters 1 and number of clusters 2 and number of clusters 3. But when the number of clusters is marked as 4 there is not much of a change. This also been approved by analysing the Elbow curve. After the number of clusters is 3 there is not much of change in the curve and we can agree that the optimum number of clusters for the given Dataset is "***Three***".

The silhoutte score method measures how tightly the observations are clustered and the average distance between clusters.

The Silhoutte Score for the given Dataset is "**0.40072705527512986".**

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

We can able to analyse that we have formed three clusters for the given Dataset.

The bank can able to give promotional offers for the clusters where there is no defaulters and the credit value is high and where they make advance payments. This will make the customer to use their credit card for the promotional offers and bank can also gain from that by getting the interest every month.

Also they can give promotional offers for the persons where their credit card usage is high and they are paying the amount at the correct time without getting defaulted. But there is a chance that these customers once they got promotional offers and spending in it, they can be defaulters since they are not paying the full amount every month, they are partially paying their interest just so that they will not be in defaulters.

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).

The sample of the data is shown below.

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
2995	28	CWT	Travel Agency	Yes	166.53	Online	364	256.20	Gold Plan	Americas
2996	35	C2B	Airlines	No	13.50	Online	5	54.00	Gold Plan	ASIA
2997	36	EPX	Travel Agency	No	0.00	Online	54	28.00	Customised Plan	ASIA
2998	34	C2B	Airlines	Yes	7.64	Online	39	30.55	Bronze Plan	ASIA
2999	47	JZI	Airlines	No	11.55	Online	15	33.00	Bronze Plan	ASIA

The info about the data is shown below

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Age           3000 non-null   int64
 1   Agency_Code   3000 non-null   object
 2   Type          3000 non-null   object
 3   Claimed       3000 non-null   object
 4   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object
 6   Duration      3000 non-null   int64
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object
 9   Destination   3000 non-null   object
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

As we can see in the info, there are features which is in Object Datatype. Before going to build model using this dataset we can to change the Object Datatype to Int so that the Model can understand.

For all the models which are going to be performed using this dataset, the model will not take object type as their input. So it is mandatory to change Object Datatype to Int.

```
Number of duplicate rows = 139
```

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel Agency	No	0.0	Online	2	20.0	Customised Plan	ASIA
422	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
...	...	...	...	...	...	...	...	...	...	...
2940	36	EPX	Travel Agency	No	0.0	Online	8	10.0	Cancellation Plan	ASIA
2947	36	EPX	Travel Agency	No	0.0	Online	10	28.0	Customised Plan	ASIA
2952	36	EPX	Travel Agency	No	0.0	Online	2	10.0	Cancellation Plan	ASIA
2962	36	EPX	Travel Agency	No	0.0	Online	4	20.0	Customised Plan	ASIA
2984	36	EPX	Travel Agency	No	0.0	Online	1	20.0	Customised Plan	ASIA

139 rows × 10 columns

As we can see there are 139 rows of duplicated data. We have to remove the duplicated rows so that we can use the Dataset to build model for more analysis.

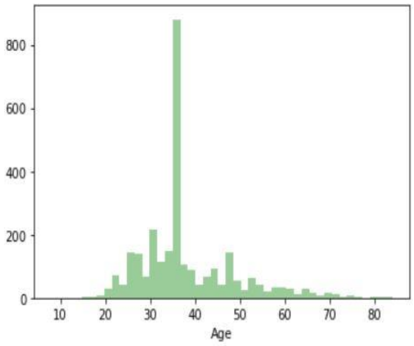
The description of the dataset is shown below.(Both Categorical and Numberical Variables are included).

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	2954	NaN	NaN	1136	2465
mean	38.091000	NaN	NaN	NaN	14.529203	NaN	70.001333	60.249913	NaN	NaN
std	10.463518	NaN	NaN	NaN	25.481455	NaN	134.053313	70.733954	NaN	NaN
min	8.000000	NaN	NaN	NaN	0.000000	NaN	-1.000000	0.000000	NaN	NaN
25%	32.000000	NaN	NaN	NaN	0.000000	NaN	11.000000	20.000000	NaN	NaN
50%	36.000000	NaN	NaN	NaN	4.630000	NaN	26.500000	33.000000	NaN	NaN
75%	42.000000	NaN	NaN	NaN	17.235000	NaN	63.000000	69.000000	NaN	NaN
max	84.000000	NaN	NaN	NaN	210.210000	NaN	4580.000000	539.000000	NaN	NaN

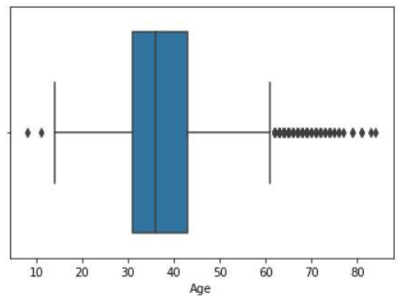
Univariate Analysis is performed for all the numerical and categorical variable and is shown below.

NUMERICAL VARIABLES:

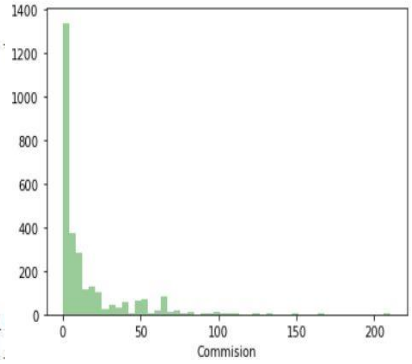
```
Description of Age
------------------------
count    2861.000000
mean       38.204124
std        10.678106
min         8.000000
25%        31.000000
50%        36.000000
75%        43.000000
max        84.000000
Name: Age, dtype: float64
```



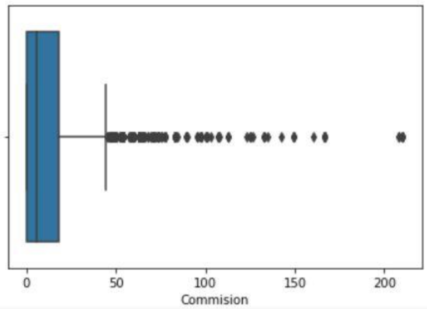
BoxPlot of Age



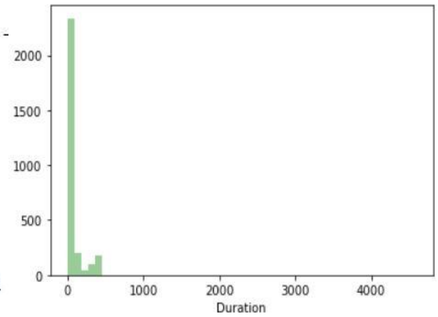

```
Description of Commision
------------------------------
count    2861.000000
mean       15.080996
std        25.826834
min         0.000000
25%         0.000000
50%         5.630000
75%        17.820000
max       210.210000
Name: Commision, dtype: float64
------------------------------
```



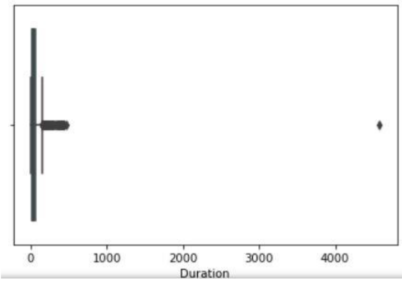
BoxPlot of Commision



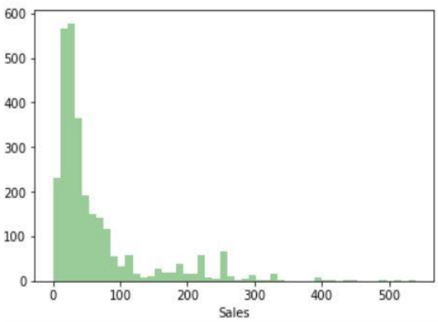

```
Description of Duration
------------------------------
count    2861.000000
mean       72.120238
std       135.977200
min        -1.000000
25%        12.000000
50%        28.000000
75%        66.000000
max      4580.000000
Name: Duration, dtype: float64
```



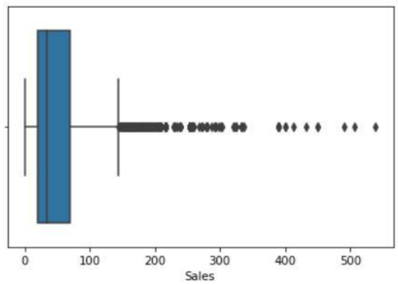
BoxPlot of Duration




```
Description of Sales
--------------------------
count    2861.000000
mean       61.757878
std        71.399740
min         0.000000
25%        20.000000
50%        33.500000
75%        69.300000
max       539.000000
Name: Sales, dtype: float64
```



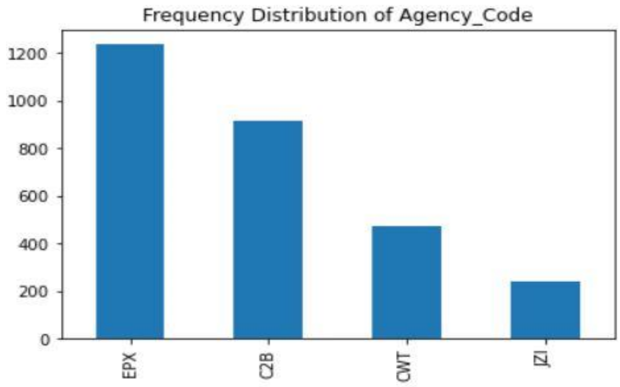
BoxPlot of Sales



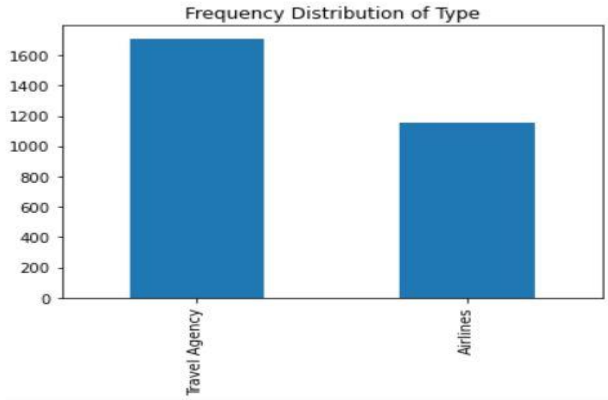
These are the Univariate Analysis for the Numerical Variable.

CATEGORICAL VARIABLE:

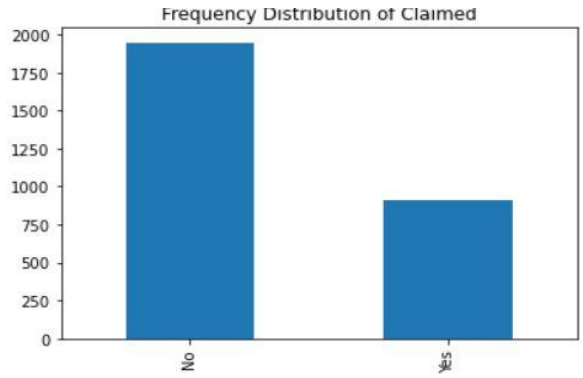
```
Details of Agency_Code
------------------------------
EPX    1238
C2B     913
CWT     471
JZI     239
Name: Agency_Code, dtype: int64
```



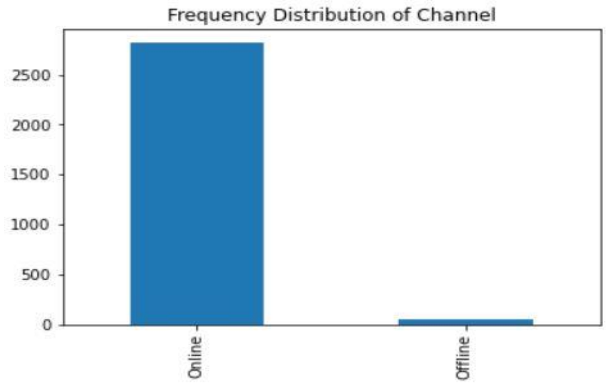

```
Details of Type
--------------------------
Travel Agency    1709
Airlines         1152
Name: Type, dtype: int64
```



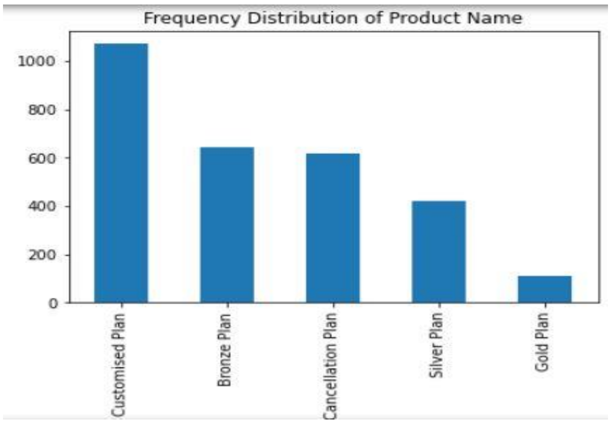

```
Details of Claimed
---------------------------
No     1947
Yes     914
Name: Claimed, dtype: int64
```



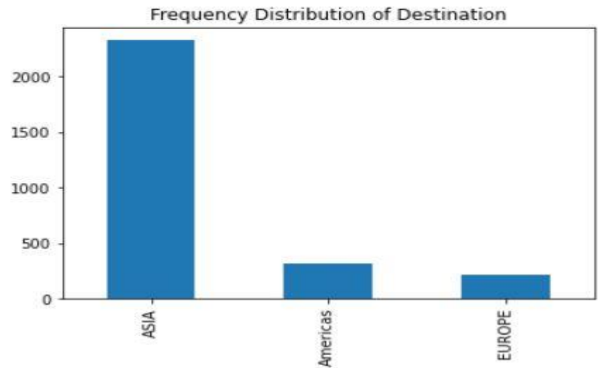
```
Details of Channel
----------------------------
Online     2815
Offline      46
Name: Channel, dtype: int64
```




```
Details of Product Name
----------------------------------
Customised Plan       1071
Bronze Plan            645
Cancellation Plan      615
Silver Plan            421
Gold Plan              109
Name: Product Name, dtype: int64
```




```
Details of Destination
------------------------------
ASIA        2327
Americas     319
EUROPE       215
Name: Destination, dtype: int64
```



INFERENCE:

- All the numerical features in the Dataset have outliers.
- The minimum insured age by the company is 8 and is maximum is 84 with an average of 36.
- The minimum commission received for tour insurance firm is 0 and the maximum is 210.10.
- The maximum duration of tour is 4580.
- There are four Agency code available which is EPX with 1238, C2B with 913, CWT with 471 and JZI with 239.

- There are two types of insrance firms which are Travel Agency with 1709 and Airlines with 1152.
- There are two Chanells available which is Online and Offline.
- There are five different type of tour insurance products available which are Customised Plan with 1071, Bronze Plan with 645, Cancellation Plan with 615, Silver Plan with 421 and Gold Plan with 109.
- There are three Destination available whih are ASIA, AMERICA and EUROPE.

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

Before splitting the data we have to convert the Object Datatypes into Numerical Datatypes so that the model can be built.

It can converted by getting the codes of the Codes of the Features. After this the splitting can be done.

The Training and Testing data is split in a ratio of 70% and 30% with a random state 1

CART: The CART model is build using "**DecisionTreeClassifier**" with the criterion as **gini** and the best parameters has been found out by using GridSearchCV and is shown below.

```
{'max_depth': 8, 'min_samples_leaf': 20, 'min_samples_split': 45}
```

Once the model is built we have to fit the Model with the training data to extract the information.

RANDOM FOREST: The Random Forest in build using the "**RandomForestClassifier**" and the best parameters is found out the GridSearchCV which is displayed below.

```
{'max_depth': 9,
 'max_features': 8,
 'min_samples_leaf': 25,
 'min_samples_split': 75,
 'n_estimators': 501}
```

The model is then fit and trained by the Training Dataset so that we can able to check the accuracy of the model by using the Test Data.

ANN: Before building the model for ANN we have to scale the data which is mandatory. This is done by using the StandardScaler which will use Z-Score to scale the data.

Once the data is scaled then the model is build using the "**MPLClassifer**" and the model is fit and trained with Training data. The best parameters of the model is established by using "**GridSearchCV**" and it is shown below.

```
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100, 100),
 'max_iter': 10000,
 'solver': 'adam',
 'tol': 0.1}
```

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: The CART model is build and it is trained by Training Data. Once the model is trained then we can use the testing data to test the accuracy of prediction of the model.

The classification report of the model for the training data is shown below.

	precision	recall	f1-score	support
0	0.82	0.88	0.85	1359
1	0.70	0.59	0.64	643
accuracy			0.79	2002
macro avg	0.76	0.74	0.75	2002
weighted avg	0.78	0.79	0.78	2002

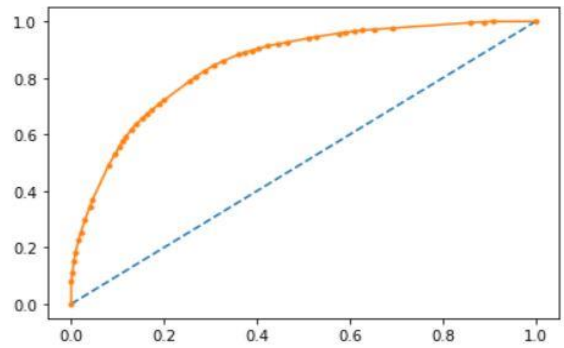
The classification report of the model for test data is shown below.

	precision	recall	f1-score	support
0	0.80	0.86	0.83	588
1	0.63	0.53	0.58	271
accuracy			0.75	859
macro avg	0.72	0.69	0.70	859
weighted avg	0.75	0.75	0.75	859

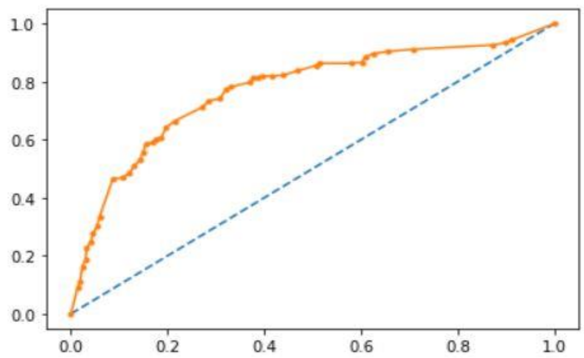
Here we can able to see that the F1-Score is higher for 0 which conveys '*No*'. The precision, recall is lower in test data compared to training data.

The AUC - Score for the Training Data is 0.849 and the AUC – Score for the testing data is 0.771 and the curve is shown below.

AUC: 0.849



AUC: 0.771



The Confusuion Matrix for the Training Data is shown below.

```
array([[1199,  160],
       [ 262,  381]], dtype=int64)
```

The confusion matrix for the Test Data is shown below.

```
array([[504,  84],
       [127, 144]], dtype=int64)
```

The accuracy of the Training Data is **0.78921** and the accuracy of the Testing data is **0.75436.**

RANDOM FOREST

The random forest model is trained using the training data. Once the model is trained we can use the testing data to predict the accuracy of the model.

The Classification report for the Training Data is shown below.

	precision	recall	f1-score	support
0	0.82	0.88	0.85	1359
1	0.70	0.59	0.64	643
accuracy			0.79	2002
macro avg	0.76	0.74	0.75	2002
weighted avg	0.78	0.79	0.78	2002

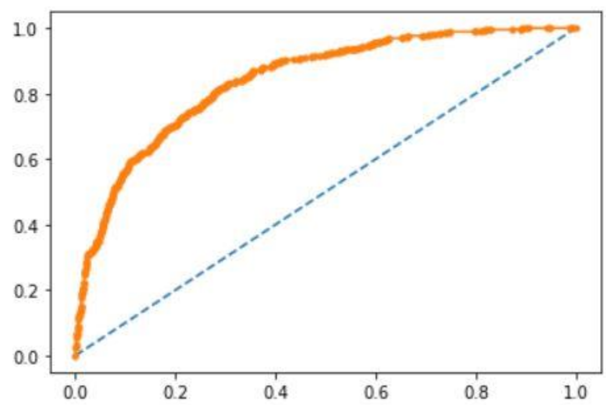
The classification report for the testing data is shown below.

	precision	recall	f1-score	support
0	0.80	0.86	0.83	588
1	0.63	0.53	0.58	271
accuracy			0.75	859
macro avg	0.72	0.69	0.70	859
weighted avg	0.75	0.75	0.75	859

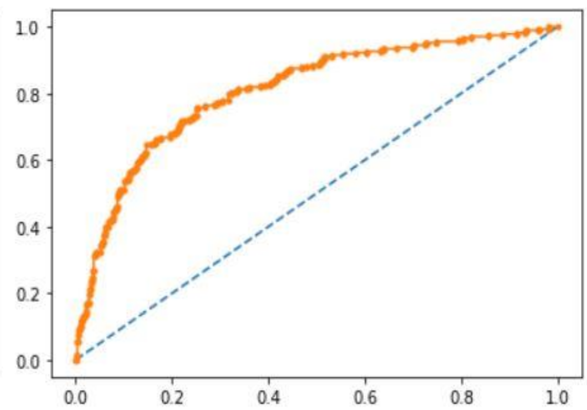
The precision, recall is almost same for both training data and testing data for '0'. The F1-Score for the training data for '0' is 0.85 and for '1' is 0.64 whereas the F1-Score for the testing data for '0' is 0.83 and '1' is 0.58.

The AUC Score is found and the AUC ROC curve for the training and testing data is shown below.

AUC: 0.843



AUC: 0.810



The confusion matrix for the training data is

```
array([[1199,  160],
       [ 262,  381]], dtype=int64)
```

The confusion matrix for the test data is

```
array([[504,  84],
       [127, 144]], dtype=int64)
```

The Accuracy of the model for training data is **0.79220** and the accuracy of the model for the test data is **0.78230.**

ANN:

The MPL Classifier is build only after scaling the data. Once the data is scaled then the model is built and trained using the training data.

The classification report of the model in training data is

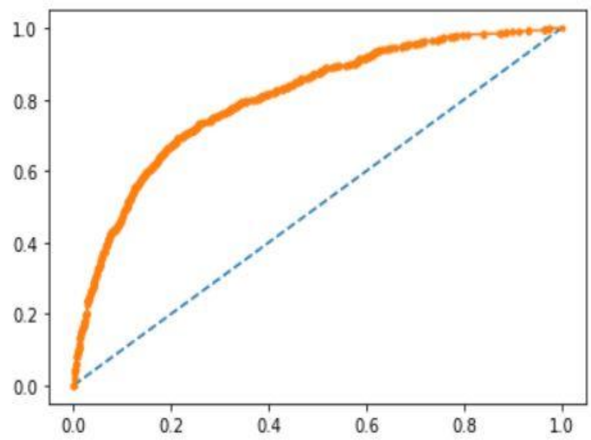
	precision	recall	f1-score	support
0	0.81	0.87	0.84	1359
1	0.67	0.57	0.61	643
accuracy			0.77	2002
macro avg	0.74	0.72	0.72	2002
weighted avg	0.76	0.77	0.76	2002

The classification report of the model on test data is

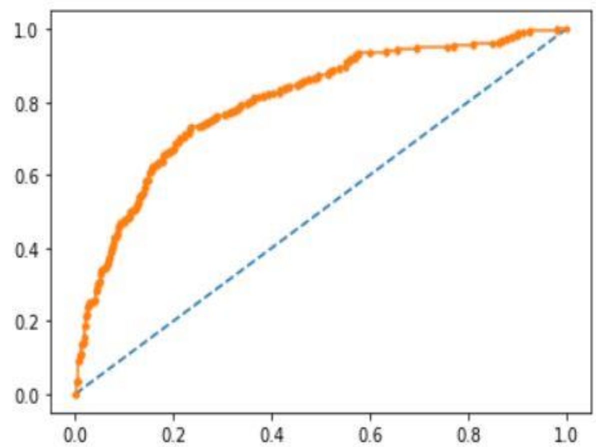
	precision	recall	f1-score	support
0	0.81	0.86	0.83	588
1	0.65	0.56	0.60	271
accuracy			0.76	859
macro avg	0.73	0.71	0.72	859
weighted avg	0.76	0.76	0.76	859

The AUC curve is build for the model using both training data and testing data and the result is shown below.

AUC: 0.803



AUC: 0.801



The confusion matrix of the model for testing data is

```
array([[1176,  183],
       [ 278,  365]], dtype=int64)
```

The confusion matrix of the model for testing data is

```
array([[506,  82],
       [120, 151]], dtype=int64)
```

The accureacy of the model for training data is **0.76973** and the accuracy of the model for testing data is **0.76484.**

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

	RECALL	F1-SCORE	PRECISION
CART	0.88	0.85	0.82
	0.59	0.64	0.70
RANDOM FOREST	0.88	0.85	0.82
	0.59	0.64	0.70
ANN	0.87	0.84	0.81
	0.57	0.61	0.67

The above table is combination of all the models along with their recall, F1 Score and Precision for the trained model.

The CART model and RANDOM FOREST have the same recall, precision and the F1 score.

	RECALL	F1-SCORE	PRECISION
CART	0.86	0.83	0.80
	0.53	0.58	0.63
RANDOM FOREST	0.86	0.83	0.80
	0.53	0.58	0.63
ANN	0.86	0.83	0.81
	0.56	0.60	0.65

The above table is a combination of recall, precision and F1 score of the test data model.

The main objective that we have to look for the tour company is facing higher claim frequency. So we have to look for persons who have already claimed and yet claiming again and we have to look for persons who are not claimed but conveying that they have claimed.

By using the confusion matrix we can analyse that the The TP are the persons who have already claimed and TN are the persons who have not claimed, FN are the persons who have already claimed but yet claiming again and FP are the persons who have not claimed but marked as claimed.

In out scenario the FN are the reasons for the Insurance company for facing higher claim frequency.

In this Dataset FN are the main score which we have to look for.

So we have to see the Sesnitivity/Recall score so that if the error is reduced then the company will not be facing higher claim frequency.

So the best model to look for is the model where the FN are less. The best model to predict this is **ARTIFICIAL NEURAL NETWORK** where the FN are less.

2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

- The insurance firm should first collect the record properly so that there will not be any duplicates.
- The insurance company should update the record as soon as the person who is claiming the insurance and once it’s approved. In this way there will not be any person who can claim their insurance twice for the same problem.
- Also they would also set a campaign for all the persons who have their insurance not claimed for more numberof days so that these will automatically gets disquallified if they are claiming after a long time.