Data mining project

—

Mandira Roy
Great learning

Content

Problem 1:

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).

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

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

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.

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

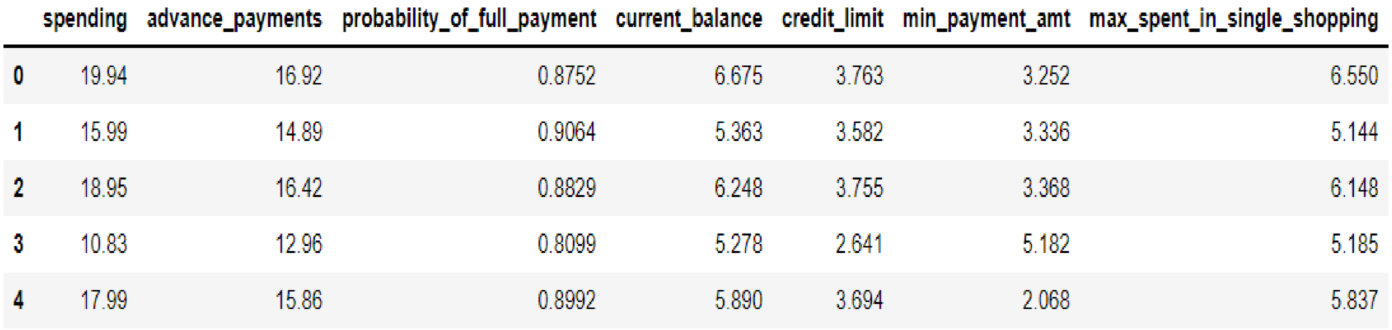
Dataset for Problem 1: bank_marketing_part1_Data.csv

Data Dictionary for Market Segmentation:

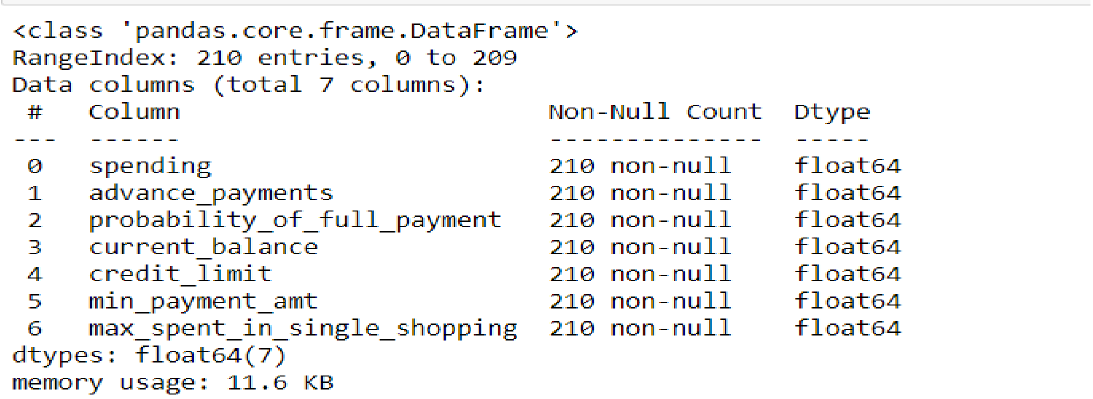
1. spending: Amount spent by the customer per month (in 1000s)
2. advance_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability_of_full_payment: Probability of payment done in full by the customer to the bank
4. current_balance: Balance amount left in the account to make purchases (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)

Answer:

1.1. Before moving ahead with the exploration of data and clustering of customers we will take a look at the data.



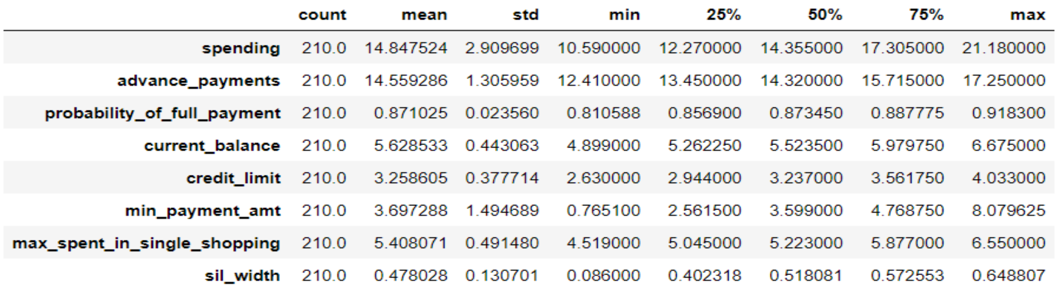
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182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



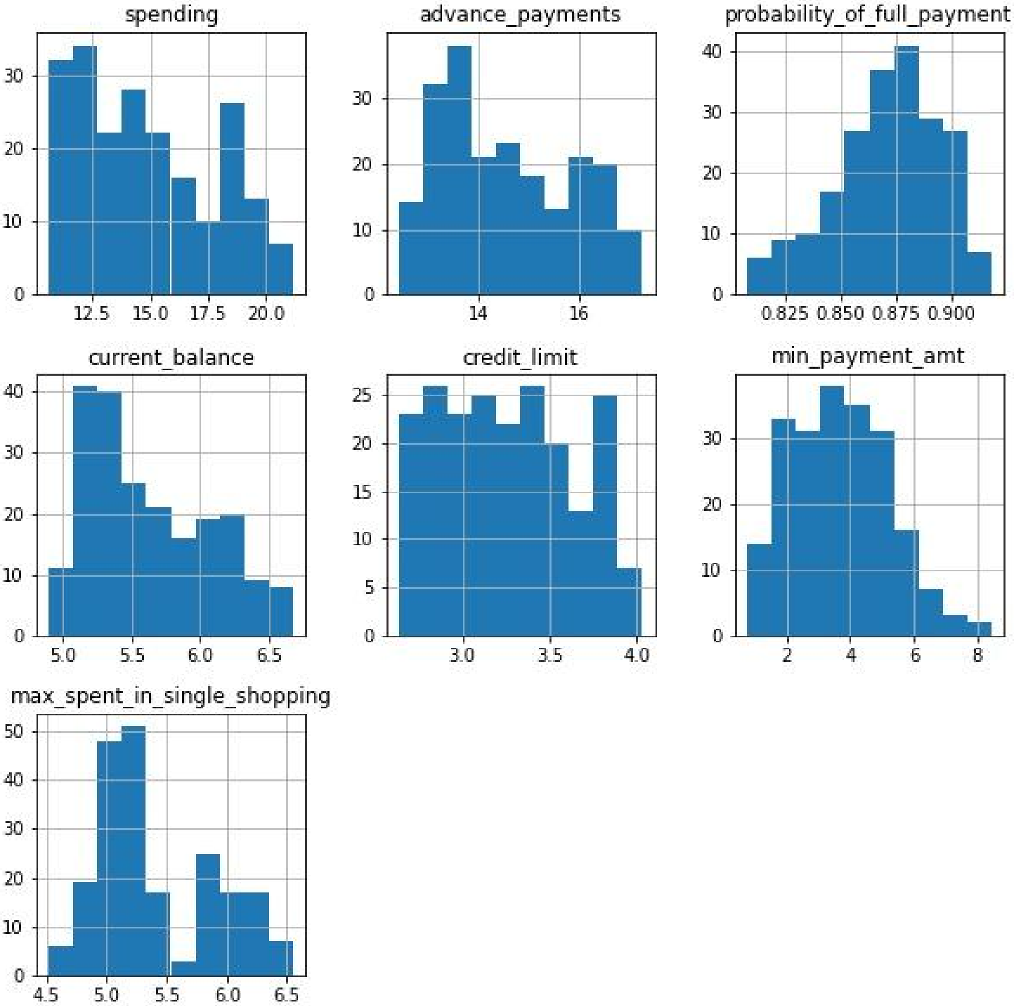
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              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
```

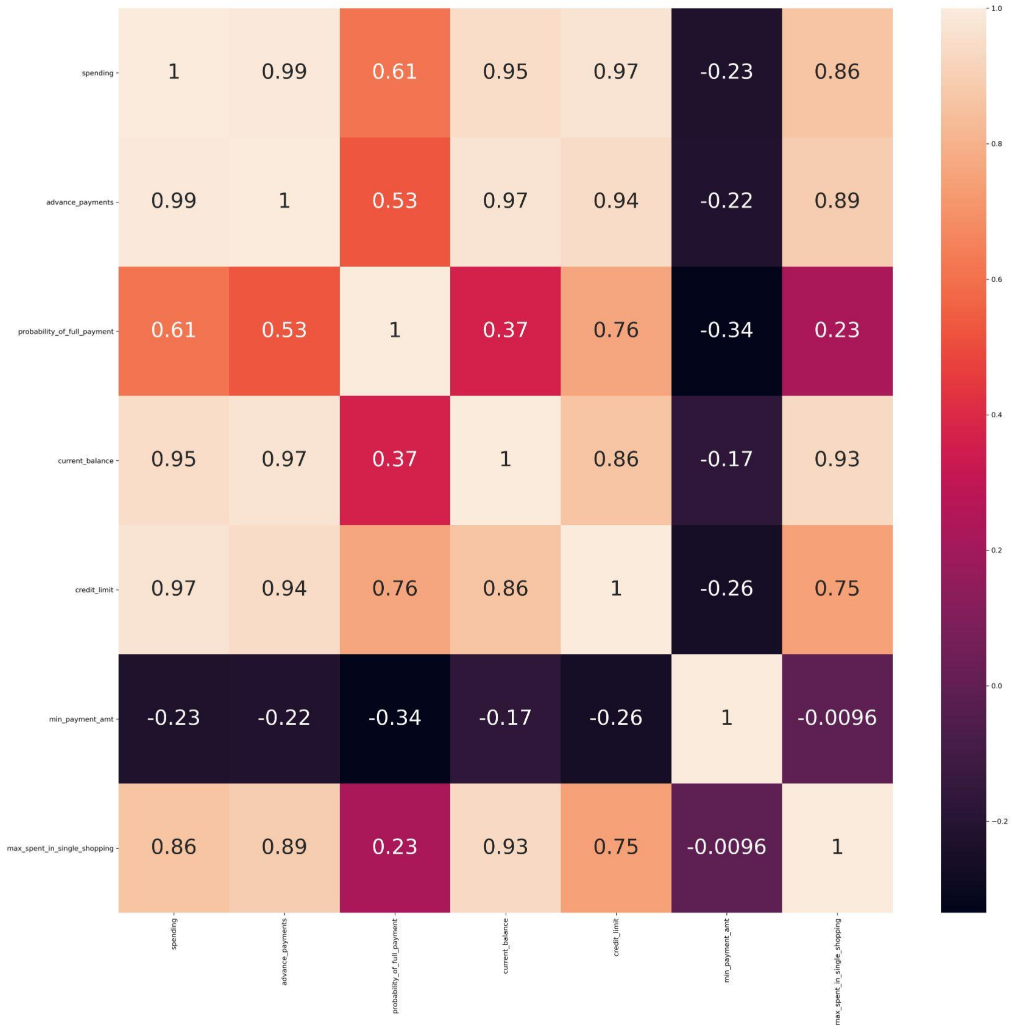
Inference: Here we can see that the data consist of 210 rows and 7 columns with no null or duplicate values and with data type float64.

Univariate analysis:



	count	mean	std	min	25%	50%	75%	max
spending	210.0	14.847524	2.909699	10.590000	12.270000	14.355000	17.305000	21.180000
advance_payments	210.0	14.559286	1.305959	12.410000	13.450000	14.320000	15.715000	17.250000
probability_of_full_payment	210.0	0.871025	0.023560	0.810588	0.856900	0.873450	0.887775	0.918300
current_balance	210.0	5.628533	0.443063	4.899000	5.262250	5.523500	5.979750	6.675000
credit_limit	210.0	3.258605	0.377714	2.630000	2.944000	3.237000	3.561750	4.033000
min_payment_amt	210.0	3.697288	1.494689	0.765100	2.561500	3.599000	4.768750	8.079625
max_spent_in_single_shopping	210.0	5.408071	0.491480	4.519000	5.045000	5.223000	5.877000	6.550000
sil_width	210.0	0.478028	0.130701	0.086000	0.402318	0.518081	0.572553	0.648807





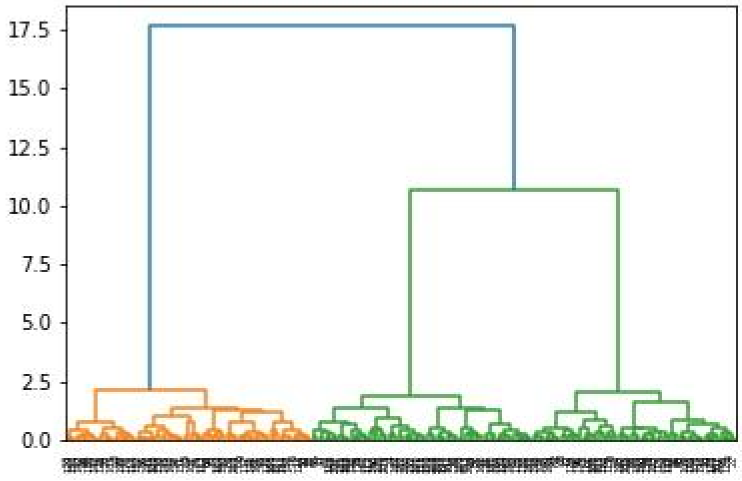
Inference:As we can notine in the heat map the highest 3 correlation is between spending vs advance_payment(0.99), spending vs credit_limit(0.97) and advance_payment vs current_balance(0.97),these are positive correlations.The top 3 negative correlations are between pobability_of_full_payment vs min_payment_amt(-0.34), credit_limit vs min_payment_amt(-0.26) and min_payment_amt vs spending(-0.23).

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

answer:Yes in this case the scaling is necessary as the mean,min,max value are't overlapping and there is a little difference in the scale of the data so here it is necessary.As the ranges of some features are not overlapping for example the standard deviation of advance_payments is 1.30 and probability_of_full_payment is 0.02 which varies widely as well as the central tendency,mean,medians varies at a significant rate. Here we will use the Min-Max scaling techniques as we would like to keep the variance that's clearly visible in the data for better modeling.

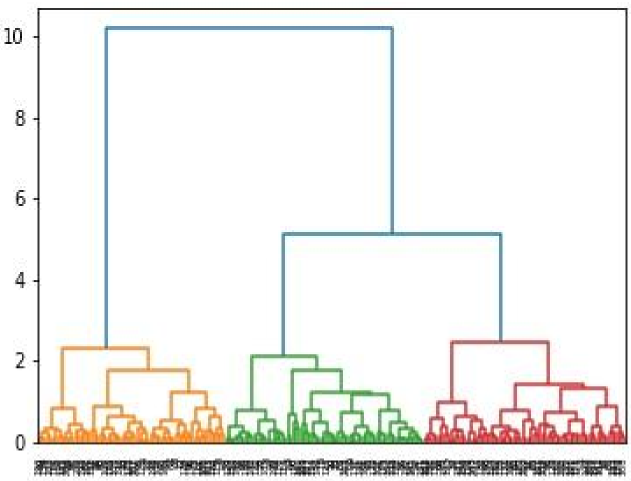
1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

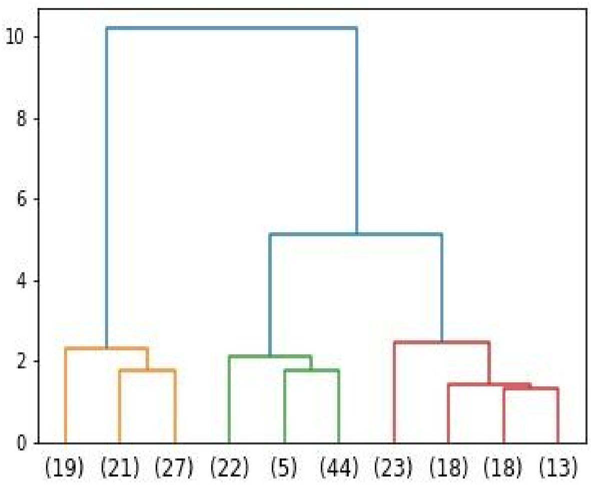
Answer:The choice of linkage and the distance metric totally depends upon the analyst .The analyst must thoroughly inspect the data ,understand the distribution of the data. For this analysis we are using the Ward linkage and Euclidean distance metric as they are the most widely used techniques by the analysts .Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function.In mathematics, the Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem, therefore occasionally being called the Pythagorean distance.

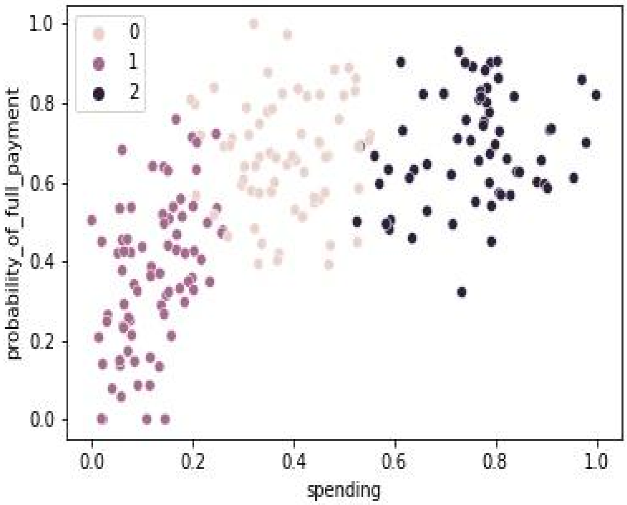


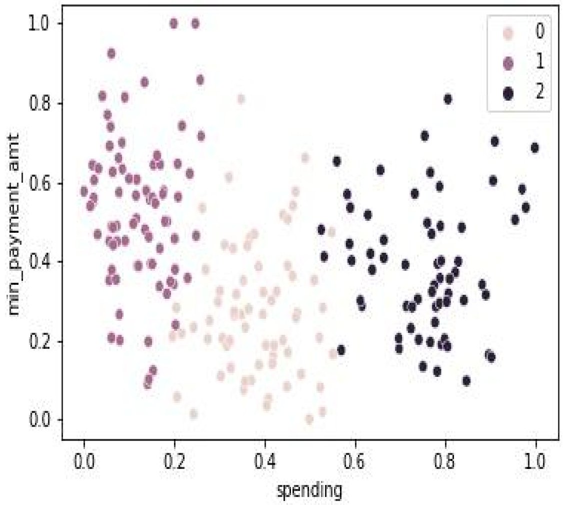
As we know how we can determine the optimal number of clusters from the dendrogram by seeing the horizontal and vertical line intersection . This is the figure where the color_threshold is 8

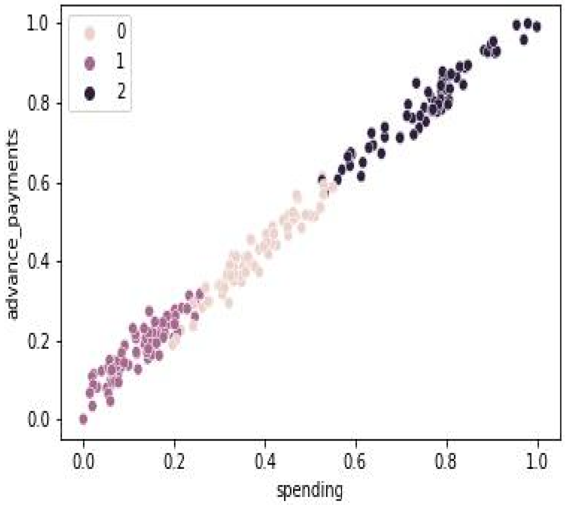
But in the next figure the color_threshold is defined as 3 i have tested color threshold 4 its giving the same result so we are using the color threshold 3 as we are getting a 3 clusters which is more preferably optimal than 2 clusters


10
8
6
4
2
0


10
8
6
4
2
0
(19)
(21)
(27)
(22)
(5)
(44)
(23)
(18)
(18)
(13)


0
1
2
probability_of_full_payment
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
spending

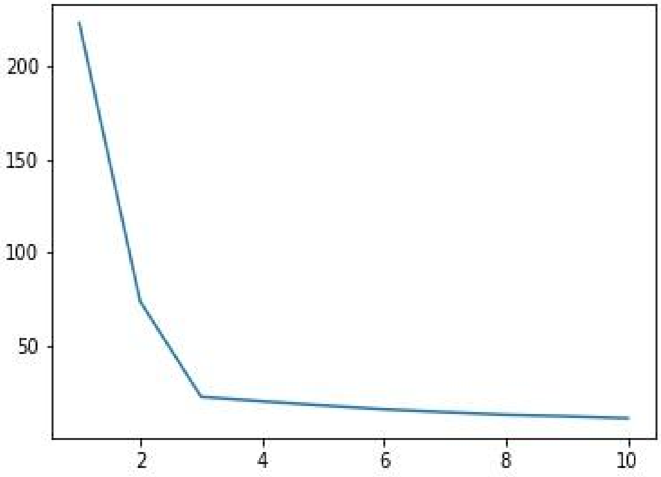

0
1
2
min_payment_amt
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
spending


0
1
2
advance_payments
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
spending

Inference: We were able to achieve distinguishable clusters among all the pairs of elements present . The clusters are defined well in the scatter plot with black ,violet and off-white colour.In most of the plot there are max black spots and minimum number of violet spots.

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.

Answer:For K-Mean clustering we need to know the K value first for that we will be using the elbow curve and the silhouette score.

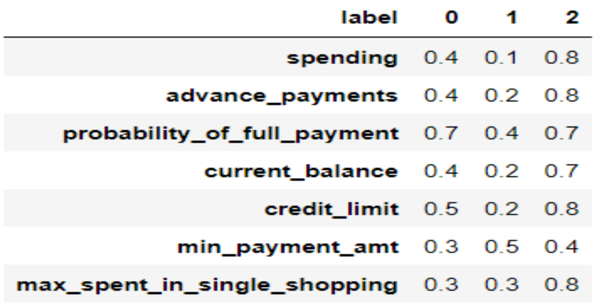


We will be looking at the elbow curve and find the k value and we will be judging that by seeing the curve . The point from which the curve is decreasing in a linear fashion that point has the value we are looking for.Here that point is 3.

We can judge it by looking at the silhouette score.As we noticed the silhouette score is close to 0.47 which quite close to one the we can say that clusters are separated,but we cant say that they are well separated and from the elbow curve we can see there is no stif decrease in the curve after 3 so the clusters 3 is the optimal value here for clustering.

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

answer:So from the cluster table we can see that the spending of customers in the cluster-1 is higher than its cluster-0 customers and then it is the cluster-2.



label	0	1	2
spending	0.4	0.1	0.8
advance_payments	0.4	0.2	0.8
probability_of_full_payment	0.7	0.4	0.7
current_balance	0.4	0.2	0.7
credit_limit	0.5	0.2	0.8
min_payment_amt	0.3	0.5	0.4
max_spent_in_single_shopping	0.3	0.3	0.8

- □ CLUSTER 0- Medium spending,medium advance_payment , highest probability_of_full_payment , medium current_balance , medium credit_limit , lowest min_payment_amt , lowest max_spent_in_single_shopping.
- □ CLUSTER 1- Lowest spending,lowest advance_payment , lowest probability_of_full_payment , lowest current_balance , lowest credit_limit , highest min_payment_amt , lowest max_spent_in_single_shopping.
- □ CLUSTER 2- Highest spending,highest advance_payment , highest probability_of_full_payment , highest current_balance , highest credit_limit , medium min_payment_amt , highest max_spent_in_single_shopping.

RECOMMENDATION: Cluster-0: We can see that cluster 2 consists of descent values. Here the customers spending is decent and one of the advance payers.and a good credit limit is maintained But we can offer them with credit points that they can avail on the next purchase on minimum mentioned amount this would promote their spendings.We can offer a loyalty card or a premium card.

Cluster-1:The highest spender in the clusters . We can increase their credit limit and provide them with some offers or discount vouchers on branded or luxurious items to increase their spending habits.We can make them an offer of minimum spending of some mentioned amount then credit points will be awarded .

Cluster-2:The least spender among the clusters but their minimum payment in single day is quite decent and this cluster consist of maximum customers so we can offer them with a some gift voucher if the minimum spending in a single day exceeds a mentioned amount or we can offer them with some rewards if they use their referrals and promote and bring more customers.

PROBLEM 2:

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).

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

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

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

Dataset for Problem 2: insurance_part2_data-1.csv

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)

7. Destination of the tour (Destination)

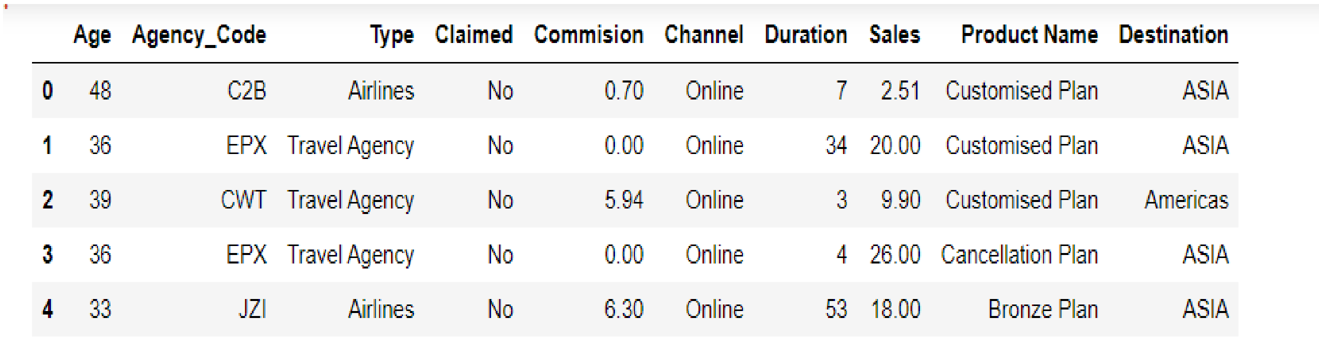
8. Amount of sales of tour insurance policies (Sales)

9. The commission received for tour insurance firm (Commission)

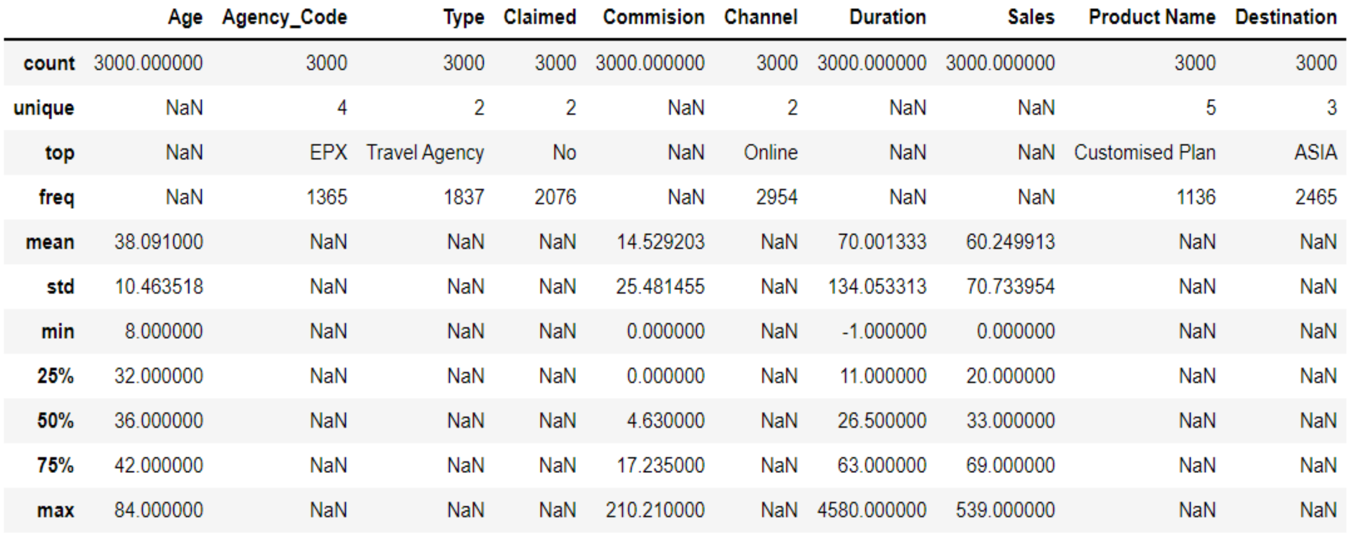
10. Age of insured (Age)

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

Answer: Before the exploratory data analysis we need to see the data and understand .For that we should we looking at data head and data description.



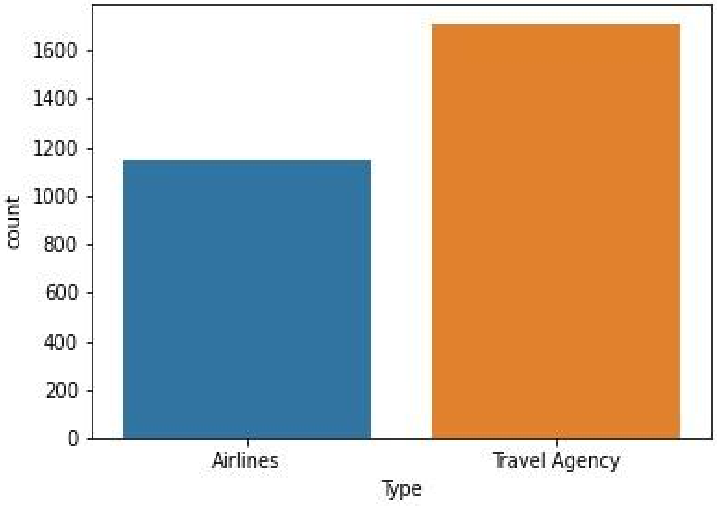
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel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

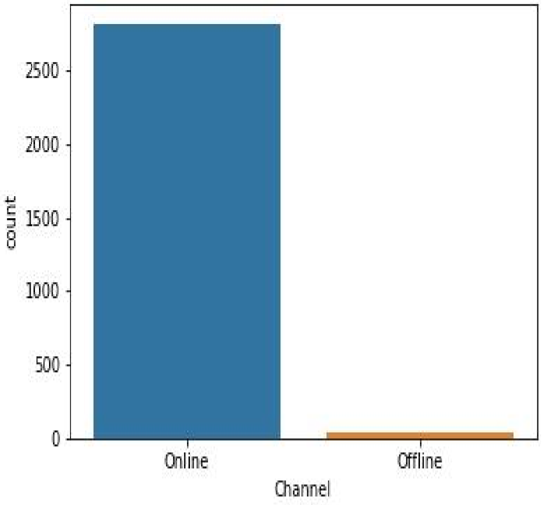


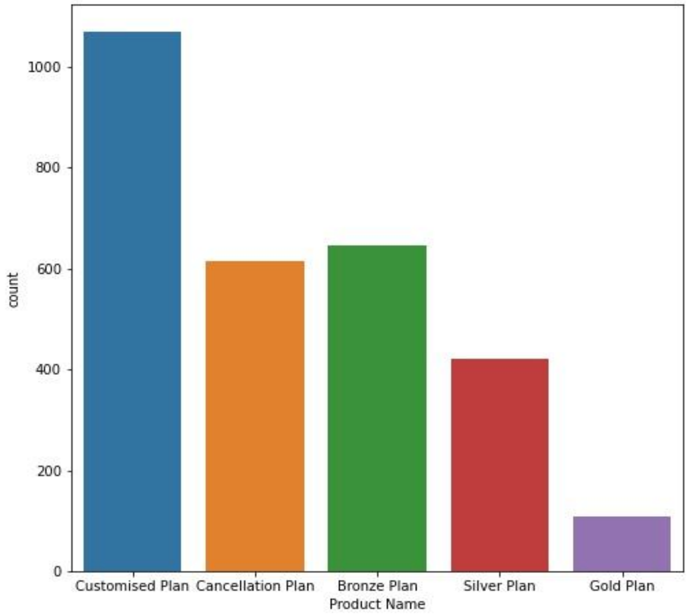
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	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

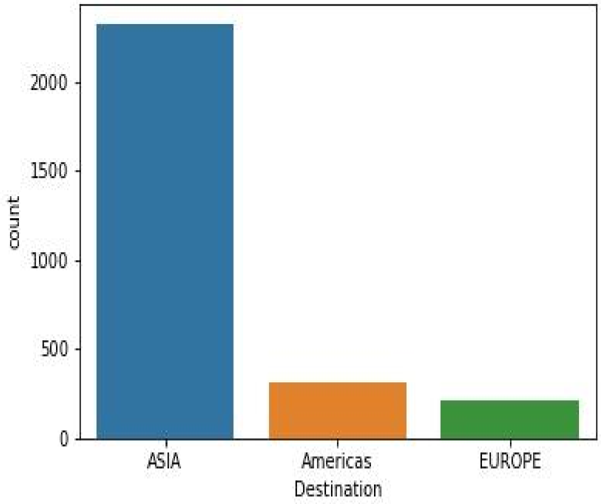
Inference: The data consist of 3000 rows and 10 columns with no null values but 139 duplicate values which will be dropped for further analysis. Agency_code,Type,Claimed, Channel,Product name, Destination are discrete variables and Age,Commision,Duration,Sales are continious variables. Claimed is the target column (i,e.,independent variable) and Agency_code,Type, Channel,Product name, Destination, Age,Commision,Duration,Sales are the dependent variables. The target variable have two categories (Yes- 924 and No-2076) i mean to say the data is a balnced data where the minority class is 45% of the majority class. (924/2076).

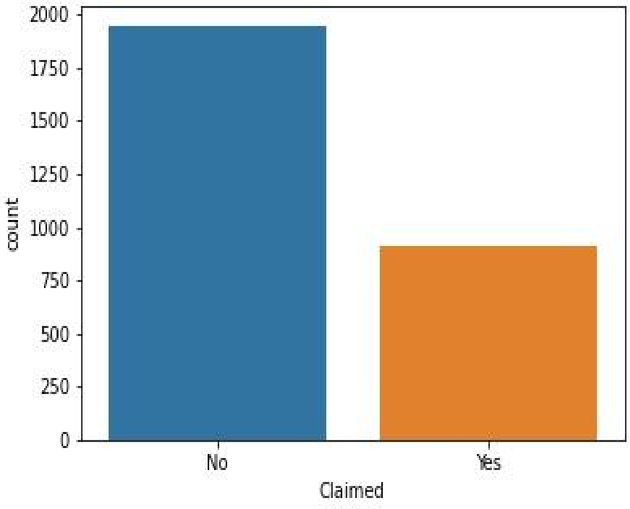
UNIVARIET ANALYSIS:











Here in the bar plots we can notice that in

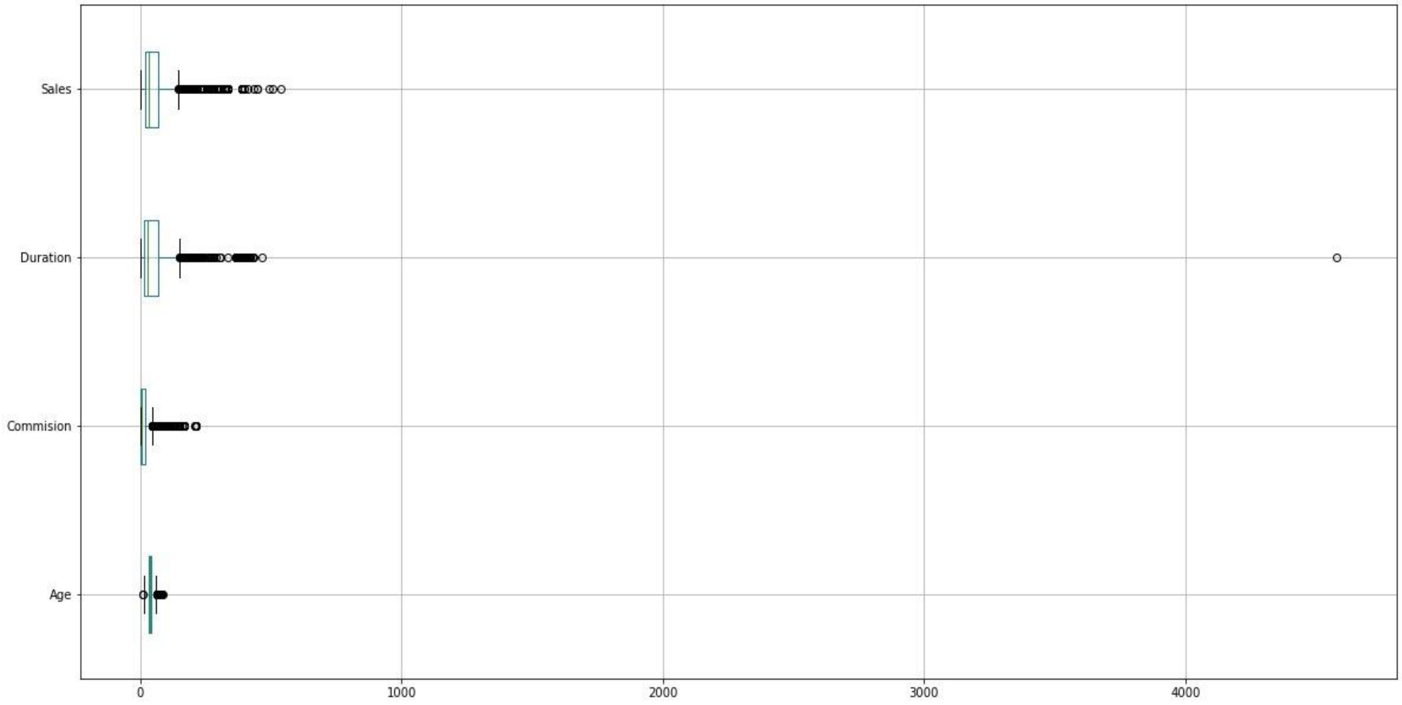
Type the Travel agency > Airlines

Channel the Online>Offline

Product Name the Customised plan>Bronze plan>Cancellation plan>Silver plan> Gold plan.

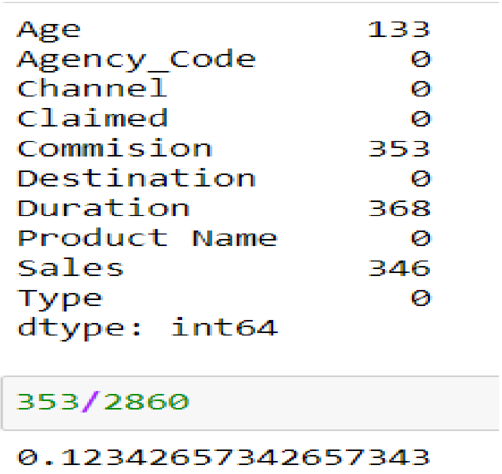
Destination the ASIA> americas> Europe

Claimed the No>Yes.



As we can see there are good number of outliers present in the dataframe but in the duration column we can find an out lier way beyond its limit and there are also some entries where duration is in 0 or -1 in those case we would normally prefer to consult with the cliet and the person who was responsible for data entry .Then we would decide to remove the annomalies or process them . Here the duation which is way beyond its limit above 1000 we are going to remove that data and the entries with 0 or -1 we are goin to replace those entries with 1. As we can see that the duration column has max number of outliers but not above the 15% limit so we are not going to treat the out liers as CART and random forest doesnt get affected by the outliers but ANN does but upto a limit i.e., if the percentage of outlier is above 15% then we will have to treat them, which is in this case 12.3%.

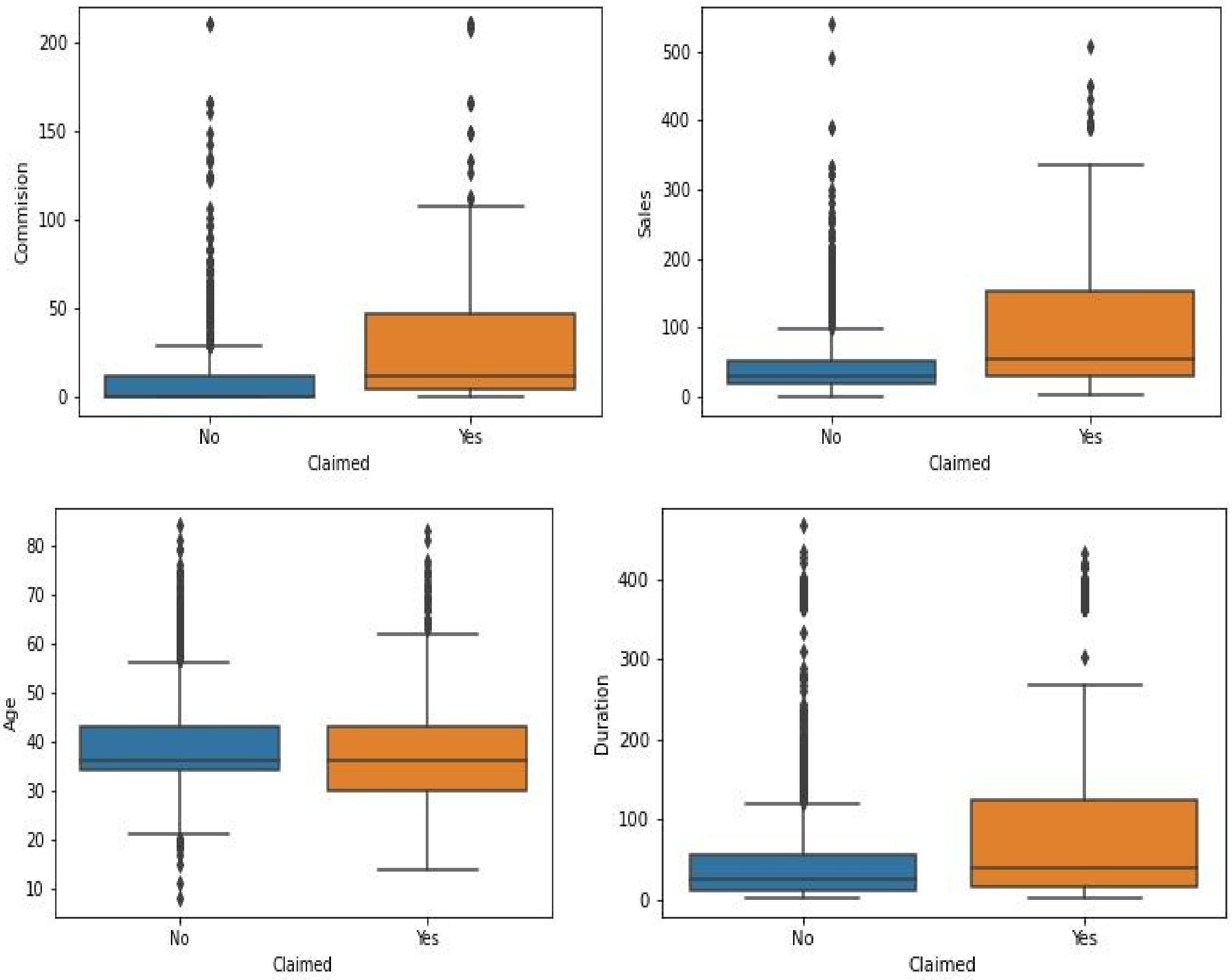
Inference:Maximum of the entries have out liers and mostly they are positively skewed .There are some entries where the sales and commisions are 0 .

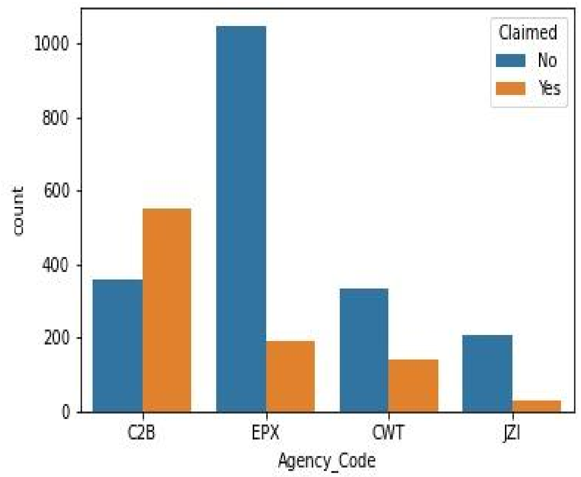

```
Age               133
Agency_Code         0
Channel             0
Claimed             0
Commision         353
Destination         0
Duration          368
Product Name        0
Sales             346
Type                0
dtype: int64
```

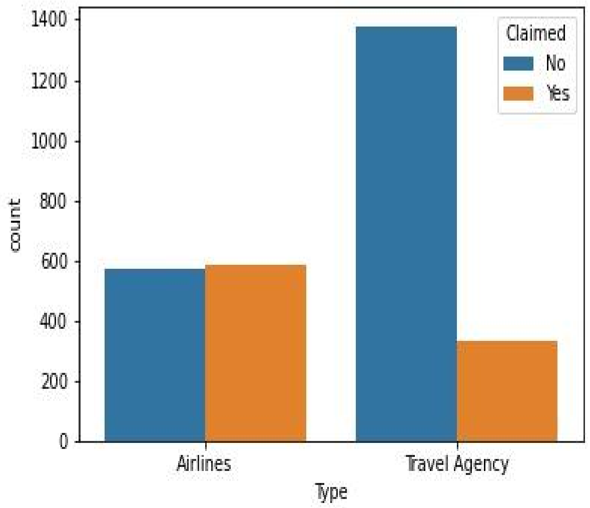
```
353/2860
```

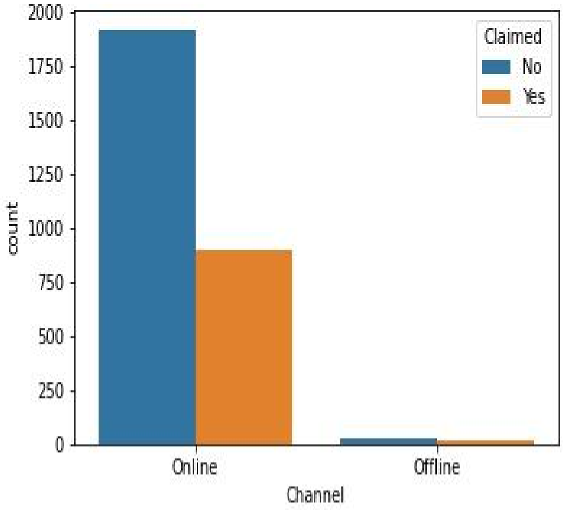
```
0.12342657342657343
```

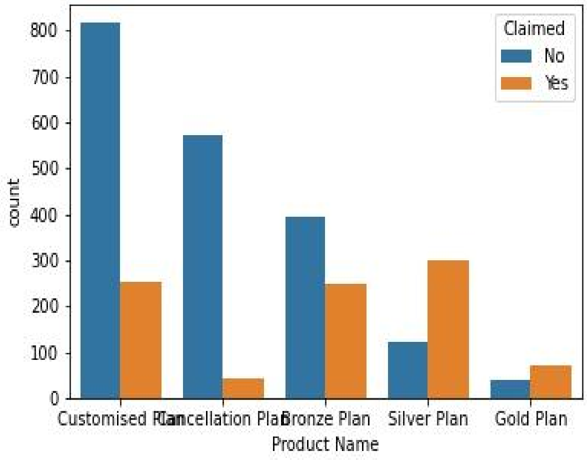
BIVARIET ANALYSIS:

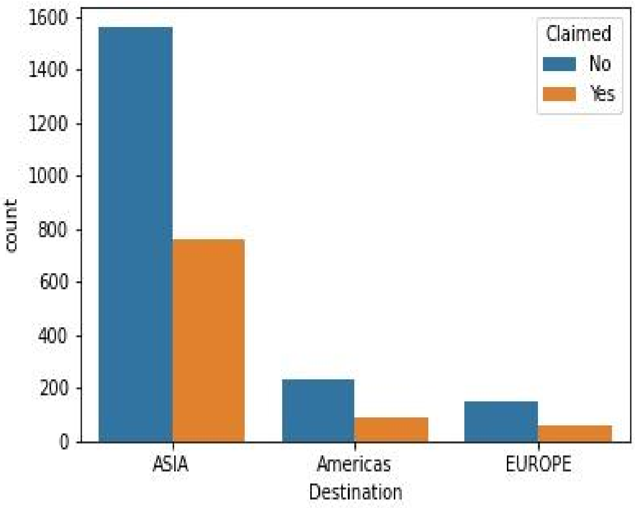






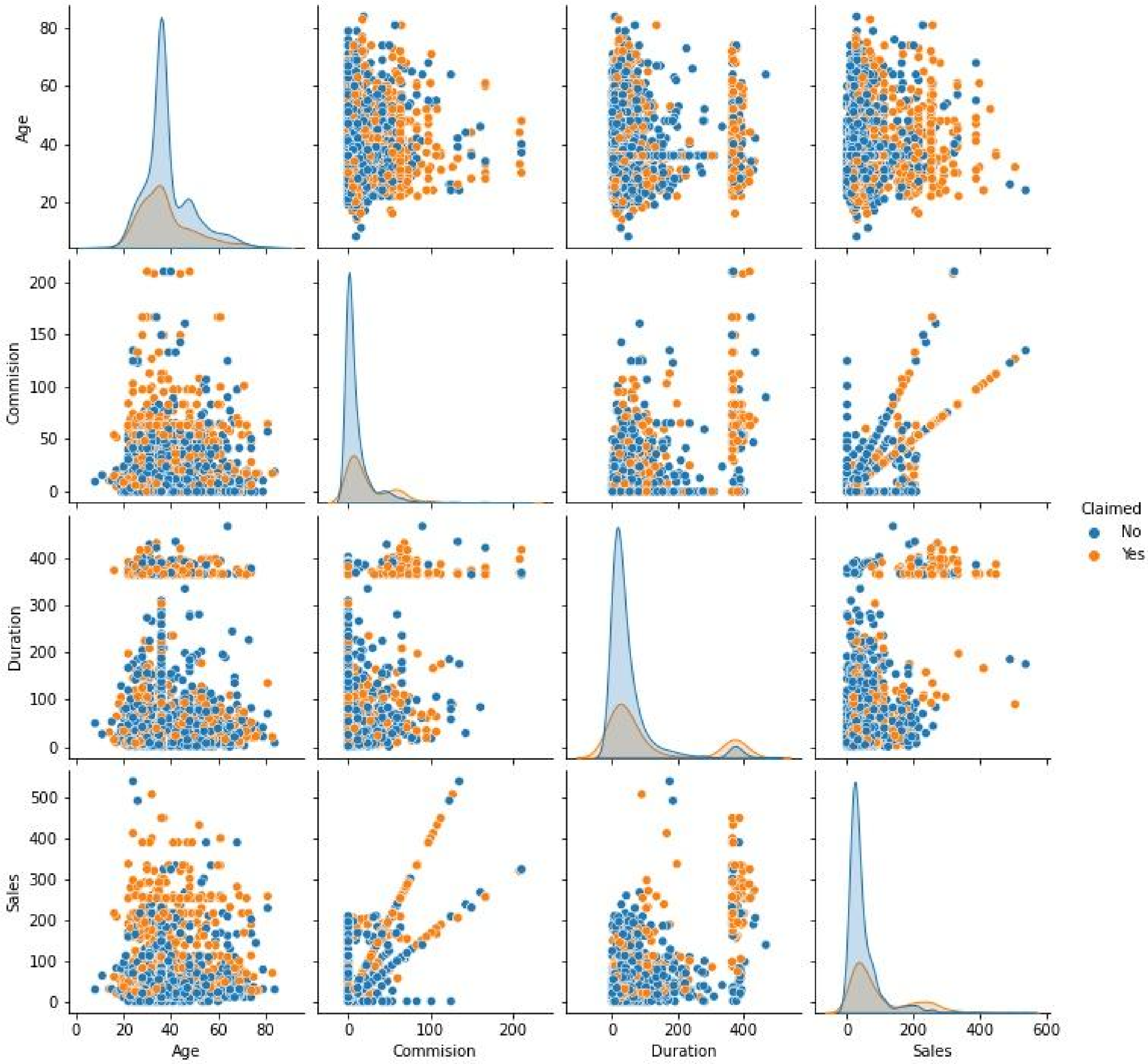


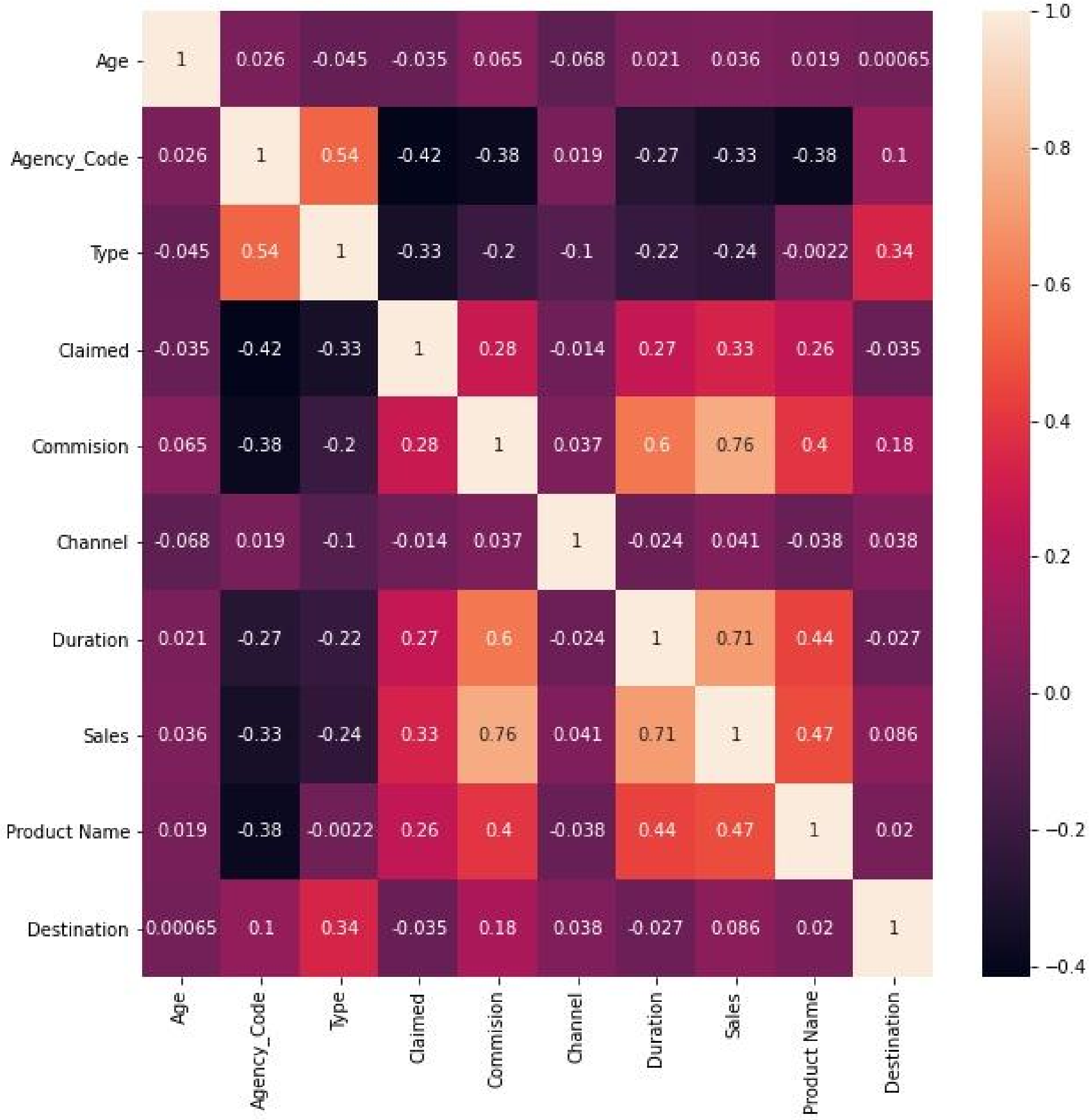




Inference: The claimed customer has similar max and min value versus those who didnot claim across the continous variables.Customers who have claimed for the insurance has similar min and max age values compare to those who didnot .Iqr for sales commision and duration has higher yes value than no .

MULTIVARIET ANALYSIS:





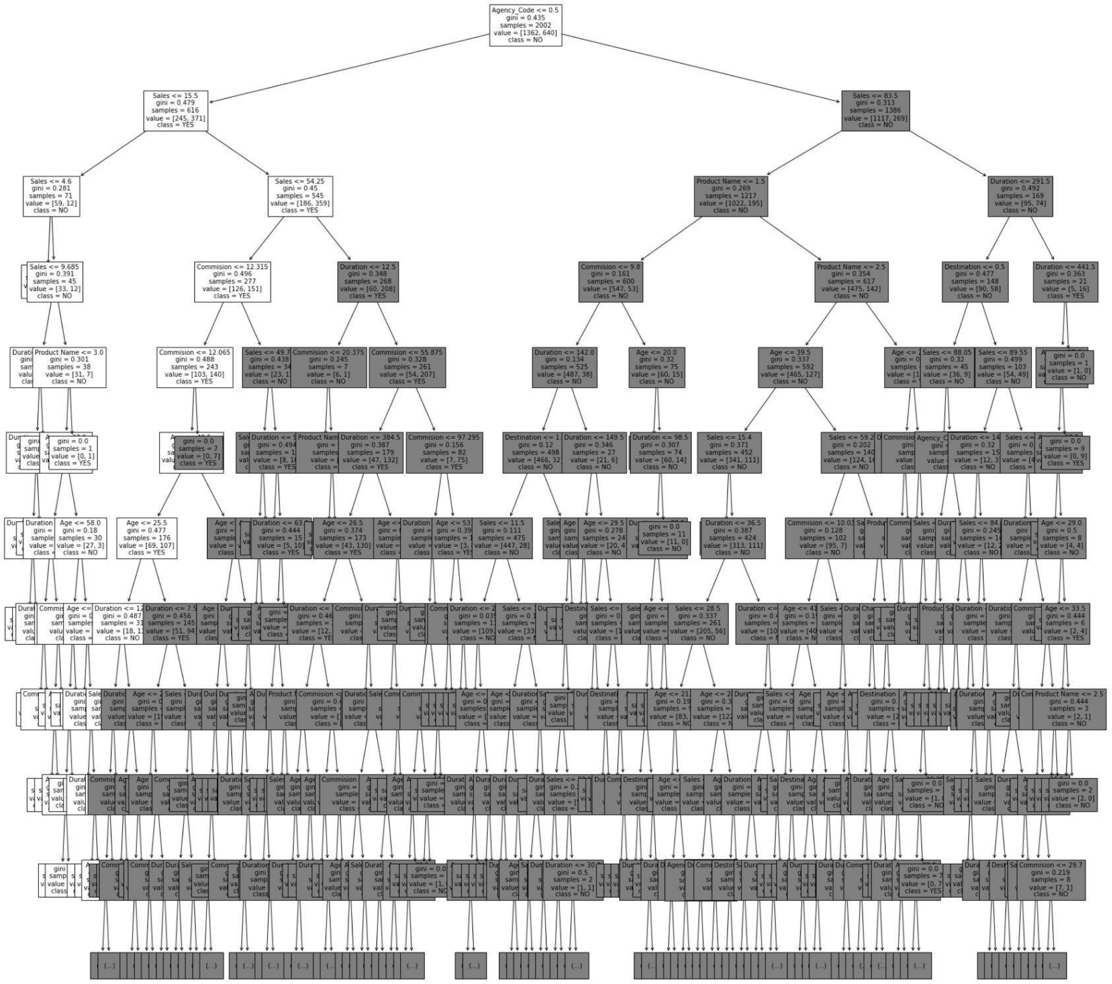
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
Age	1	0.026	-0.045	-0.035	0.065	-0.068	0.021	0.036	0.019	0.00065
Agency_Code	0.026	1	0.54	-0.42	-0.38	0.019	-0.27	-0.33	-0.38	0.1
Type	-0.045	0.54	1	-0.33	-0.2	-0.1	-0.22	-0.24	-0.0022	0.34
Claimed	-0.035	-0.42	-0.33	1	0.28	-0.014	0.27	0.33	0.26	-0.035
Commision	0.065	-0.38	-0.2	0.28	1	0.037	0.6	0.76	0.4	0.18
Channel	-0.068	0.019	-0.1	-0.014	0.037	1	-0.024	0.041	-0.038	0.038
Duration	0.021	-0.27	-0.22	0.27	0.6	-0.024	1	0.71	0.44	-0.027
Sales	0.036	-0.33	-0.24	0.33	0.76	0.041	0.71	1	0.47	0.086
Product Name	0.019	-0.38	-0.0022	0.26	0.4	-0.038	0.44	0.47	1	0.02
Destination	0.00065	0.1	0.34	-0.035	0.18	0.038	-0.027	0.086	0.02	1

INFERENCE:

There is an high correlation between commision vs sales(0.76),duration vs sales(0.71), commision vs duration(0.6). Though these correlations are not so high but these are the top three we are not getting any other insights from the figures.

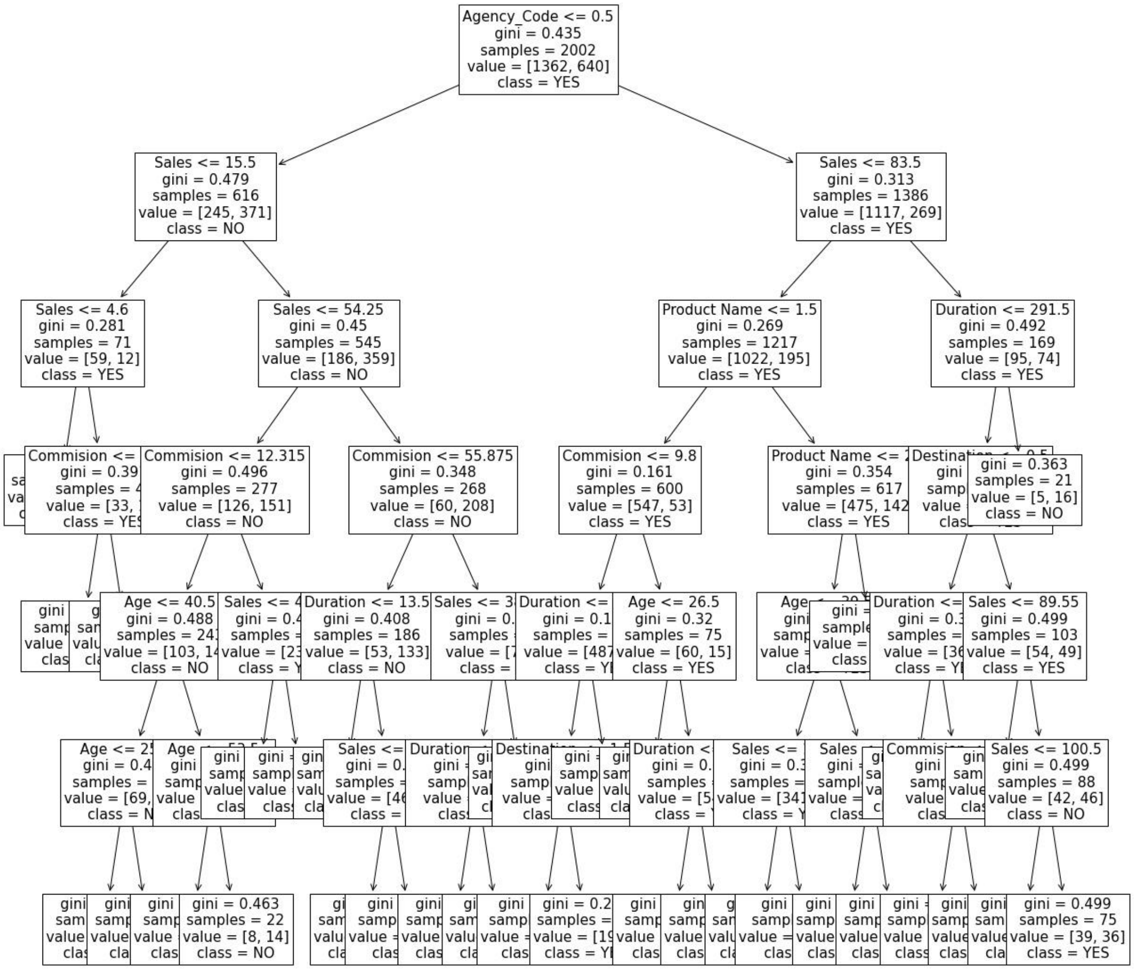
2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

Answer:



Agency_Code <= 0.5
gini = 0.435
samples = 2002
value = [1362, 640]
class = NO
Sales <= 15.5
gini = 0.479
samples = 616
value = [245, 371]
class = YES
Sales <= 83.5
gini = 0.313
samples = 1386
value = [1117, 269]
class = NO
Sales <= 4.6
gini = 0.281
samples = 71
value = [59, 12]
class = NO
Sales <= 54.25
gini = 0.45
samples = 545
value = [186, 359]
class = YES
Product Name <= 1.5
gini = 0.269
samples = 1217
value = [1022, 195]
class = NO
Duration <= 291.5
gini = 0.492
samples = 169
value = [95, 74]
class = NO
Sales <= 9.685
gini = 0.391
samples = 45
value = [33, 12]
class = NO
Commision <= 12.315
gini = 0.496
samples = 277
value = [126, 151]
class = YES
Duration <= 12.5
gini = 0.348
samples = 268
value = [60, 208]
class = YES
Commision <= 9.8
gini = 0.161
samples = 600
value = [547, 53]
class = NO
Product Name <= 2.5
gini = 0.354
samples = 617
value = [475, 142]
class = NO
Destination <= 0.5
gini = 0.477
samples = 148
value = [90, 58]
class = NO
Duration <= 441.5
gini = 0.363
samples = 21
value = [5, 16]
class = YES
Product Name <= 3.0
gini = 0.301
samples = 38
value = [31, 7]
class = NO
Commision <= 12.065
gini = 0.488
samples = 243
value = [103, 140]
class = YES
Commision <= 20.375
gini = 0.245
samples = 7
value = [6, 1]
class = NO
Commision <= 55.875
gini = 0.328
samples = 261
value = [54, 207]
class = YES
Duration <= 142.0
gini = 0.134
samples = 525
value = [487, 38]
class = NO
Age <= 20.0
gini = 0.32
samples = 75
value = [60, 15]
class = NO
Age <= 39.5
gini = 0.337
samples = 592
value = [465, 127]
class = NO
Sales <= 88.05
gini = 0.32
samples = 45
value = [36, 9]
class = NO
Sales <= 89.55
gini = 0.499
samples = 103
value = [54, 49]
class = NO
gini = 0.0
samples = 1
value = [1, 0]
class = NO
(...)

Clearly from the figure we can see that the tree has over grown and this may laed to over-fitted model. We need to prune the model and then try modeling .



Here is a better model with a some better values and it has been pruned.:I have done the modeling in the python file and in the param_grid i have tried and tested different values which were giving better results according to those values i have delected the values of parameters or i have used the default technique taught in one of the videos to use 0.1/0.2/0.3 part of the total value in min sample leaf and three times of min sample leaf is the min sample split.

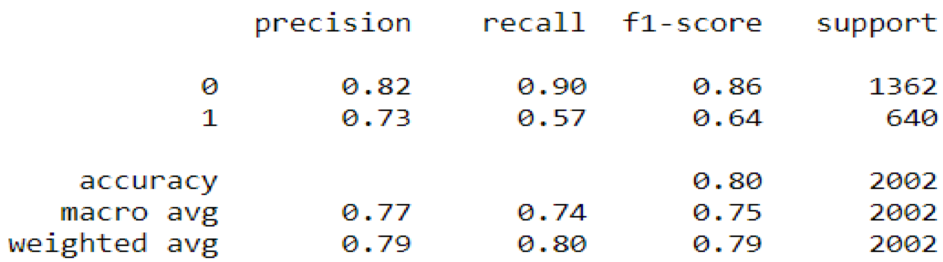
DECISION TREE:

TRAIN accuracy of- 80%

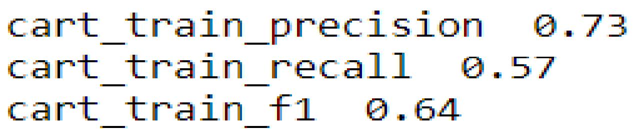
Test accuracy of- 77.27%

Train roc_auc score- 0.8384

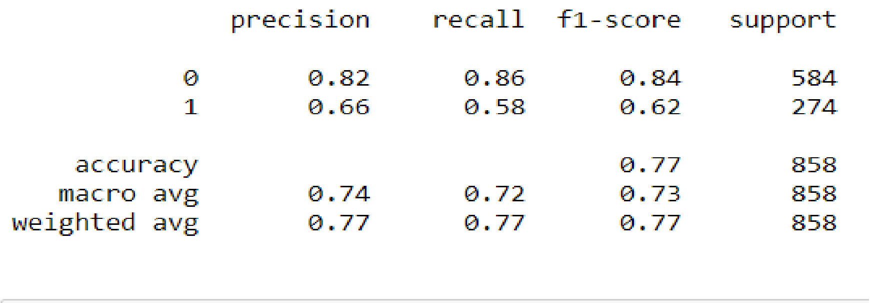
Test roc_auc score-0.805981



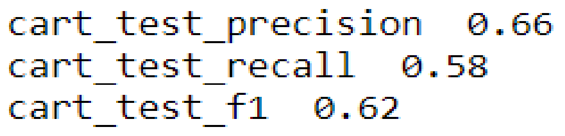
```
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1362
           1       0.73      0.57      0.64       640

    accuracy                           0.80      2002
   macro avg       0.77      0.74      0.75      2002
weighted avg       0.79      0.80      0.79      2002
```



```
cart_train_precision  0.73
cart_train_recall  0.57
cart_train_f1  0.64
```

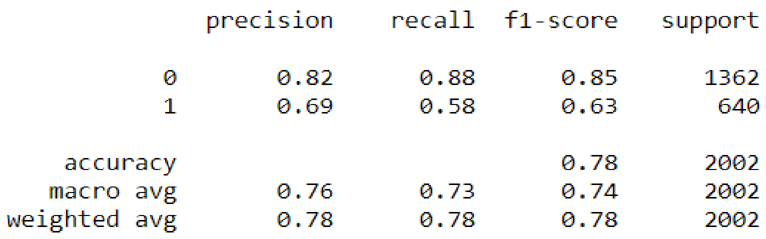


```
              precision    recall  f1-score   support

           0       0.82      0.86      0.84       584
           1       0.66      0.58      0.62       274

    accuracy                           0.77       858
   macro avg       0.74      0.72      0.73       858
weighted avg       0.77      0.77      0.77       858
```



```
cart_test_precision  0.66
cart_test_recall  0.58
cart_test_f1  0.62
```

RANDOM FOREST



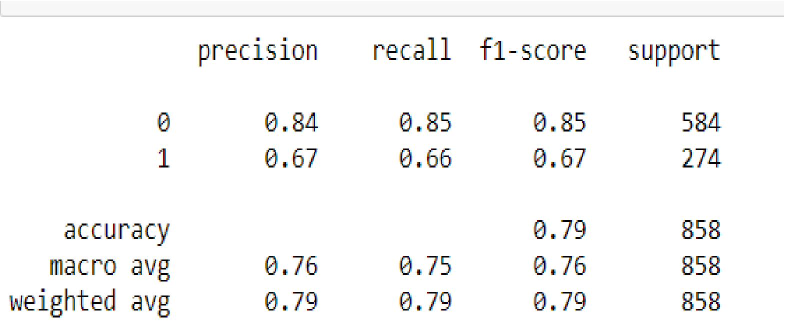
```
              precision    recall  f1-score   support

           0       0.82      0.88      0.85      1362
           1       0.69      0.58      0.63       640

    accuracy                           0.78      2002
   macro avg       0.76      0.73      0.74      2002
weighted avg       0.78      0.78      0.78      2002
```

TRAIN accuracy of- 78.42%

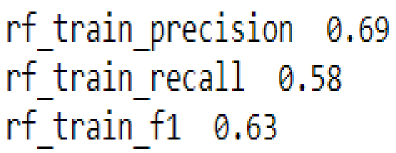
Test accuracy of- 78.9%

Train roc_auc score- 0.7308

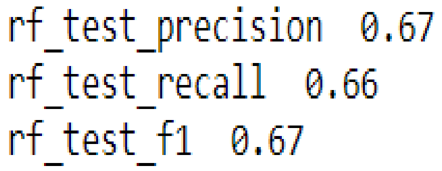
Test roc_auc score-0.7539



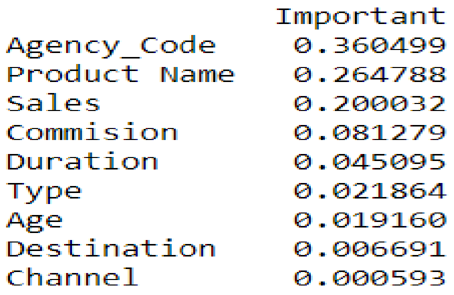
```
              precision    recall  f1-score   support

           0       0.84      0.85      0.85       584
           1       0.67      0.66      0.67       274

    accuracy                           0.79       858
   macro avg       0.76      0.75      0.76       858
weighted avg       0.79      0.79      0.79       858
```



```
rf_train_precision  0.69
rf_train_recall  0.58
rf_train_f1  0.63
```



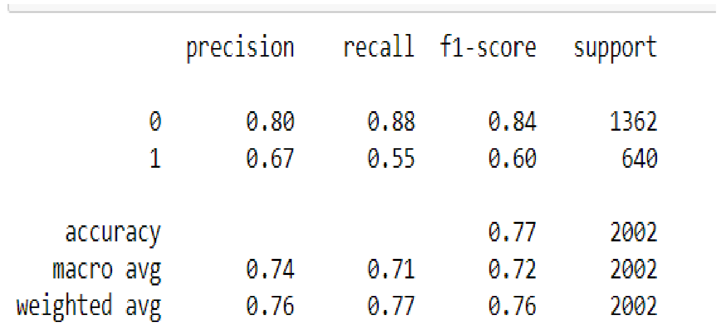
```
rf_test_precision  0.67
rf_test_recall  0.66
rf_test_f1  0.67
```



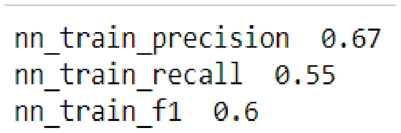
```
                Important
Agency_Code      0.360499
Product Name     0.264788
Sales            0.200032
Commision        0.081279
Duration         0.045095
Type             0.021864
Age              0.019160
Destination      0.006691
Channel          0.000593
```

INFERENCE:The most important features according to the prdiction is Agency code ,product name, sales.The least important feature is channel.The best parm grids used in the modeling is {'max_depth': 5, 'max_features': 5,'min_samples_leaf': 5,'min_samples_split': 90, 'n_estimators': 1000}

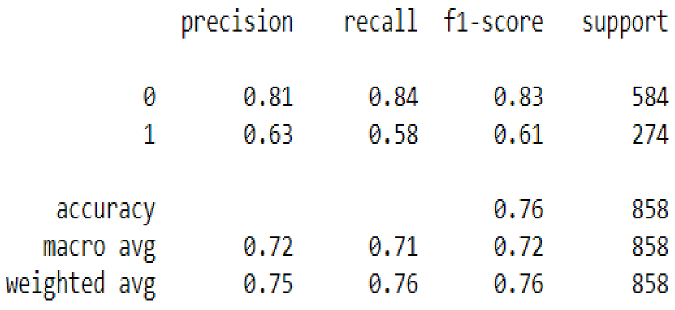
ANN MODEL



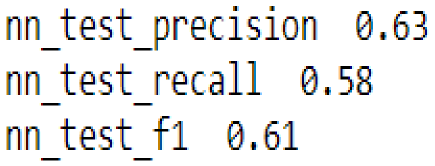
```
              precision    recall  f1-score   support

           0       0.80      0.88      0.84      1362
           1       0.67      0.55      0.60       640

    accuracy                           0.77      2002
   macro avg       0.74      0.71      0.72      2002
weighted avg       0.76      0.77      0.76      2002
```



```
nn_train_precision  0.67
nn_train_recall  0.55
nn_train_f1  0.6
```



```
              precision    recall  f1-score   support

           0       0.81      0.84      0.83       584
           1       0.63      0.58      0.61       274

    accuracy                           0.76       858
   macro avg       0.72      0.71      0.72       858
weighted avg       0.75      0.76      0.76       858
```



```
nn_test_precision  0.63
nn_test_recall  0.58
nn_test_f1  0.61
```

TRAIN accuracy of- 77%

Test roc_auc score-0.64

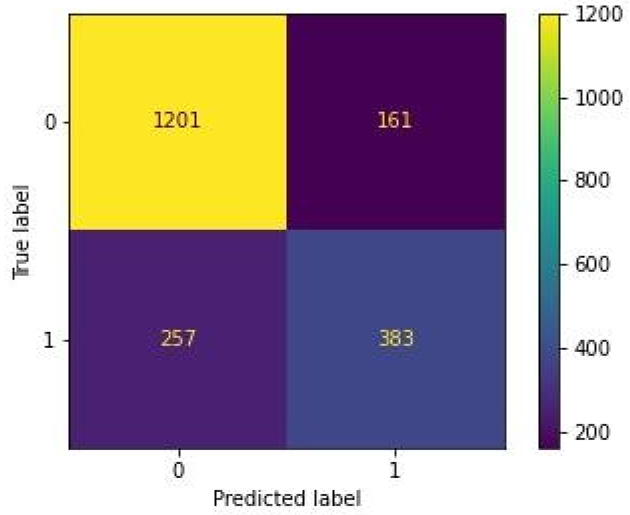
Train roc_auc score- 0.62

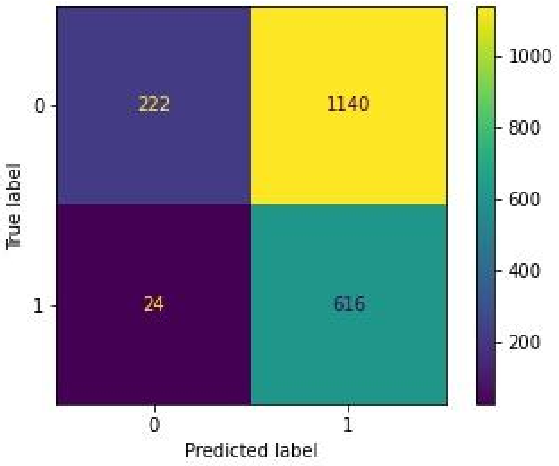
Test accuracy of- 76%

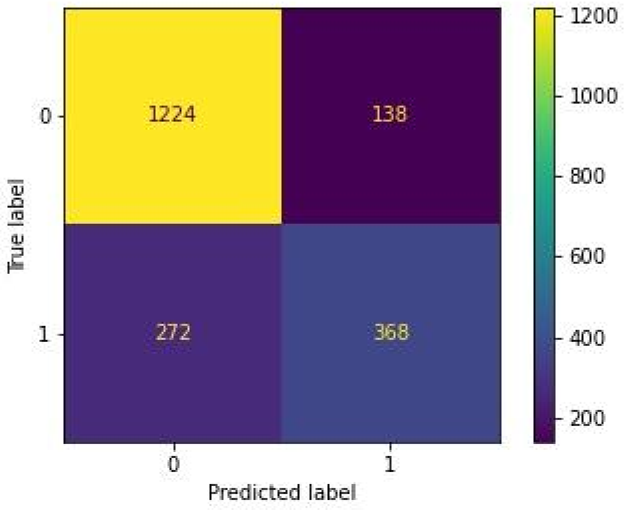
INFERENCE:

The best parm grids used is {'hidden_layer_sizes': 1000, 'max_iter': 8000, 'solver': 'adam', 'tol': 0.01}

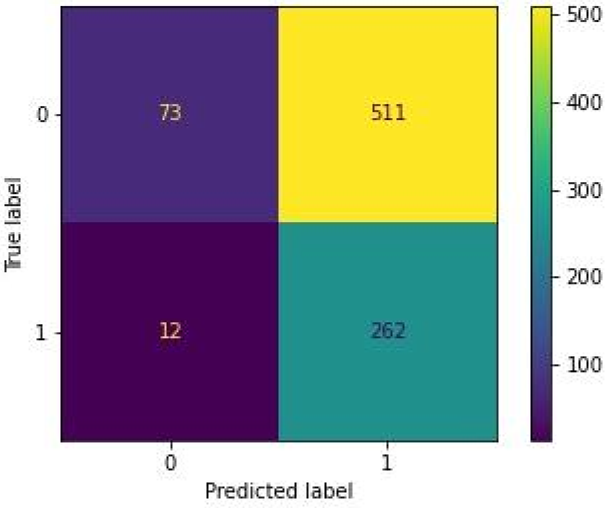
With the default activation function(relu).

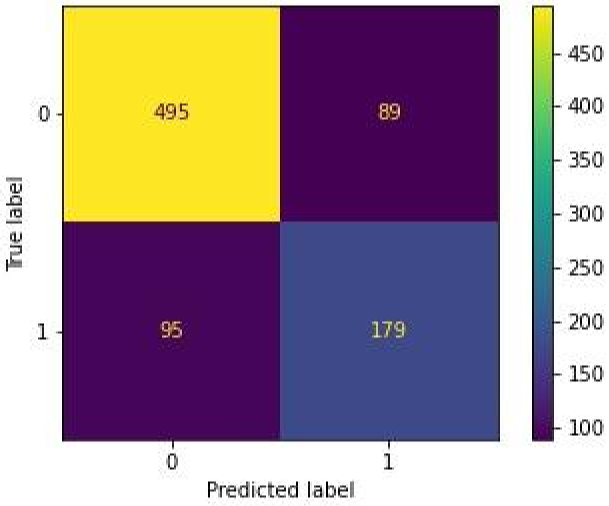


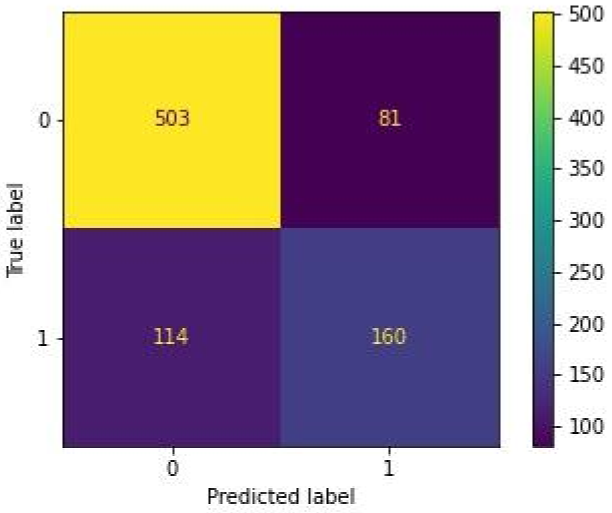




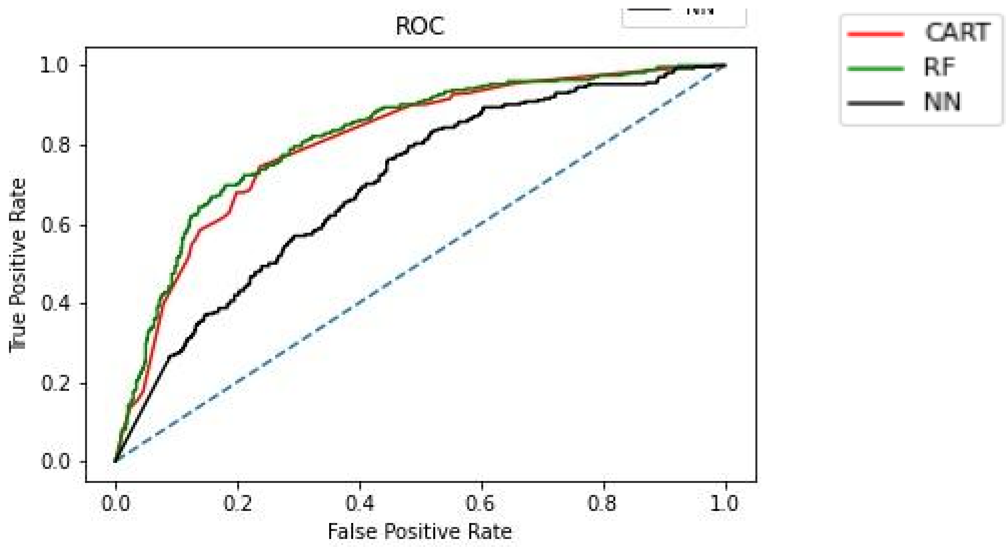
These are the confusion matrix for the train dataset random forest, ANN, CART model respectively. Next is the confusion matrix for test dataset in ANN , CART, RANDOM forest respectively

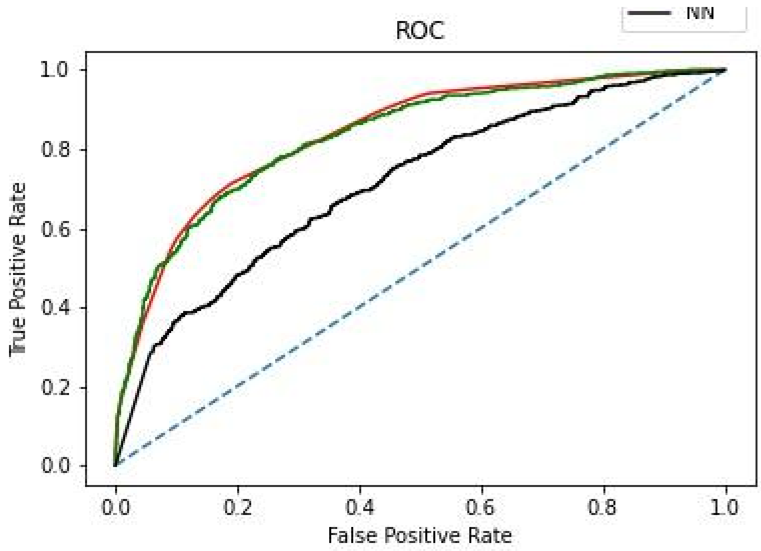


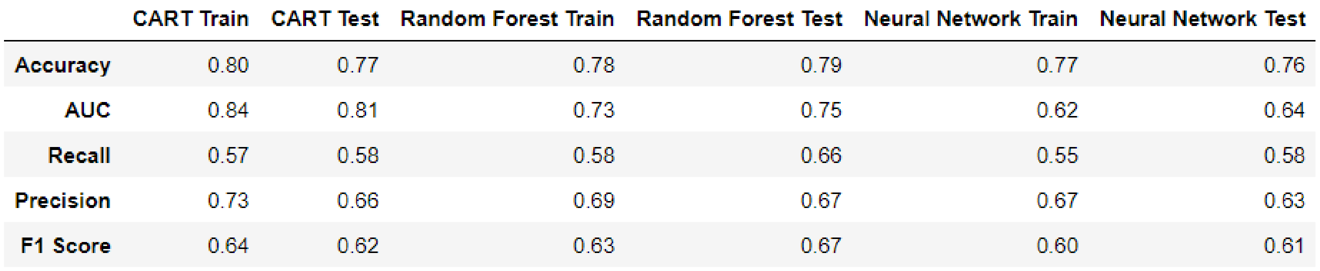




2.4 Final Model: Compare all the models and write an inference which model is best/optimized.







	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.80	0.77	0.78	0.79	0.77	0.76
AUC	0.84	0.81	0.73	0.75	0.62	0.64
Recall	0.57	0.58	0.58	0.66	0.55	0.58
Precision	0.73	0.66	0.69	0.67	0.67	0.63
F1 Score	0.64	0.62	0.63	0.67	0.60	0.61

INFERENCE:Based on the above figure we can say that the ANN RF model is performing quite good according to the accuracy score and the difference between the test metric and train metric is quite less than the CART model. The recall score is quite important in this analysis as which customer will be likely to claim the issuance . The ANN was showing a better recall score than the other models.

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

Answer: By analysing the data I think we would need more precise features and data for better analysing and for better prediction and analysing the data.We are at 80% accuracy so we need customers to plan and book tickets ,it cross sells the insurance now based on claim date pattern . Another interesting fact is that more claims and sales occur through airlines than agencies.So if we could increase the time duration for insurance claims in Airlines category then we can make a profit from that and less claims will be forwarded.From plotting we also derived a conclusion that agency_code -JZI is facing a low sales rate so i think we should give them with some marketing strategies and provide them with some products that is more popular at the market if possible.Europe as a destination has a high claim rate ,i think we need to find out the reason why is it happening .The features at importance is product names with product code -Gold plan, Silver plan, so we need to do an analysis on why these two products are more popular and why other products aren't ,more marketing and improvement in the products offerings may bring more sale.We were able to create a generalized ANN model with a recall of 0.58. So that's my conclusion and business report.

Milestones

I. Lorem ipsum

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

II. Dolor sit amet

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.