DATA MINING PROJECT-21ST AUGUST, 2022

Submitted by,

Deepa .K

Table of Contents

List of Figures:

Problem Statement 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.

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)

1.1: Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bivariate and multivariate analysis).

Solution:

The dataset has 7 columns and 210 entries with no duplicated data and all the entries are filled with float data type.

Shape: (210, 7)

Data type: Float

Length: **7** Columns

Duplication: No duplicated values in the dataset.

Missing values: No missing values in the dataset.

Summary dataset: Count, mean, std.deviation, range of values and IQR range.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

Figure 4: Summary Dataset

Observation:

- Based on the above figure, the data looks good with no missing variables.

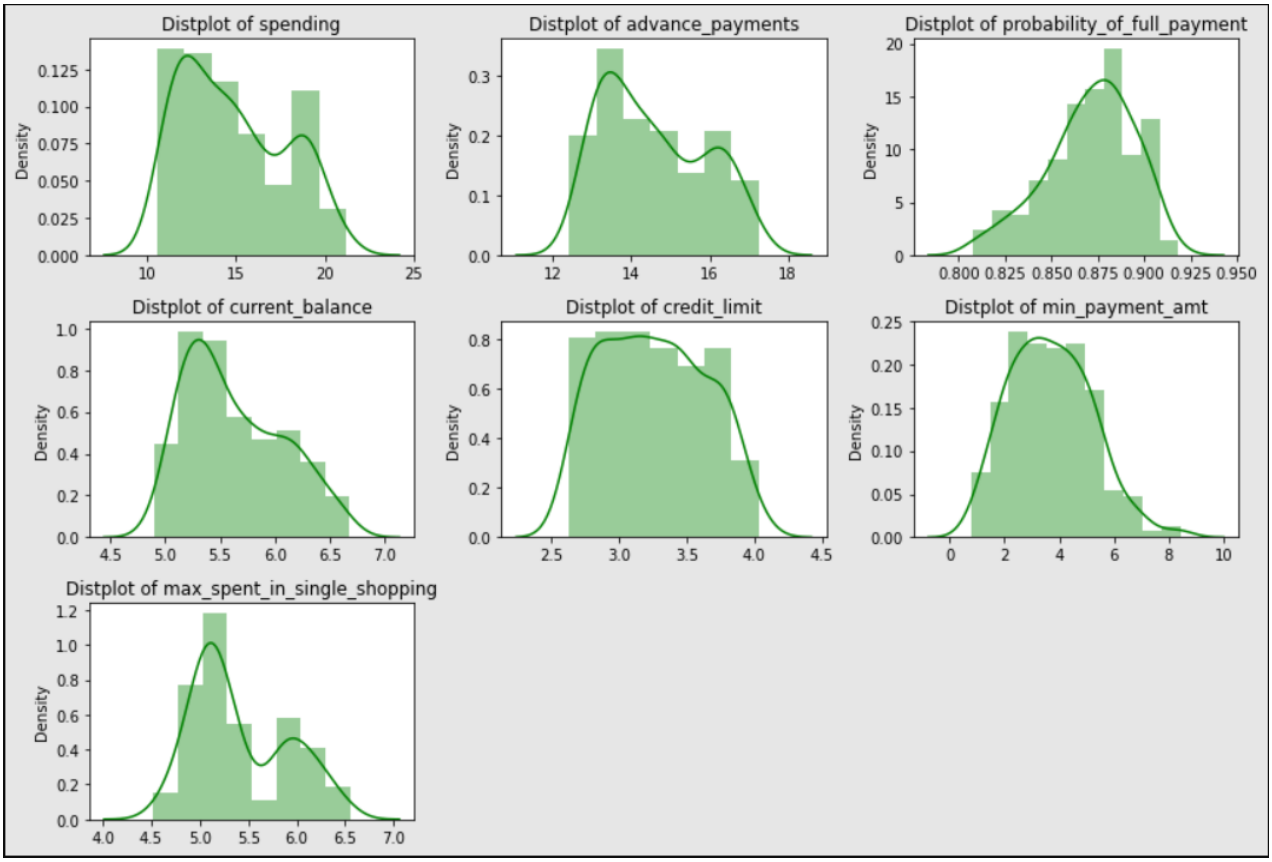

Figure 5: Distplot

Observation:

- The Distplot distribution is almost right skewed for all the variables except "Probability of full payment".

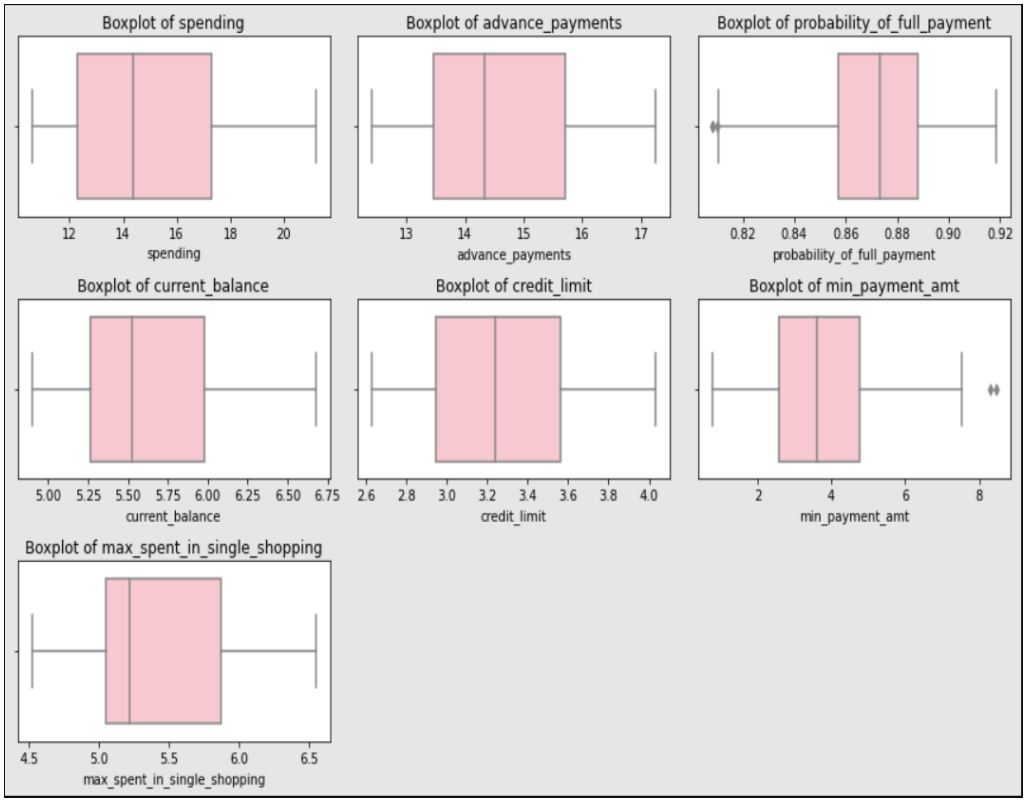

Figure 6:Boxplot

Observation:

- Spending: Lower outlier 4.71 and Upper outlier 24.85
- Advance payment: Lower outlier 10.05 and Upper outlier 19.11
- Prob. of full payment: Lower outlier 0.81 and Upper outlier 0.93
- Current balance: Lower outlier 4.18 and Upper outlier 7.05
- Credit limit: Lower outlier 2.01 and Upper outlier 4.48
- Min payment amount: Lower outlier -0.74 and Upper outlier 8.07
- Max spent in single shopping: Lower outlier 3.79 and Upper outlier 7.12
- The outliers are present in the highlighted attributes and can be replaced with median values to retain them within the dataset.

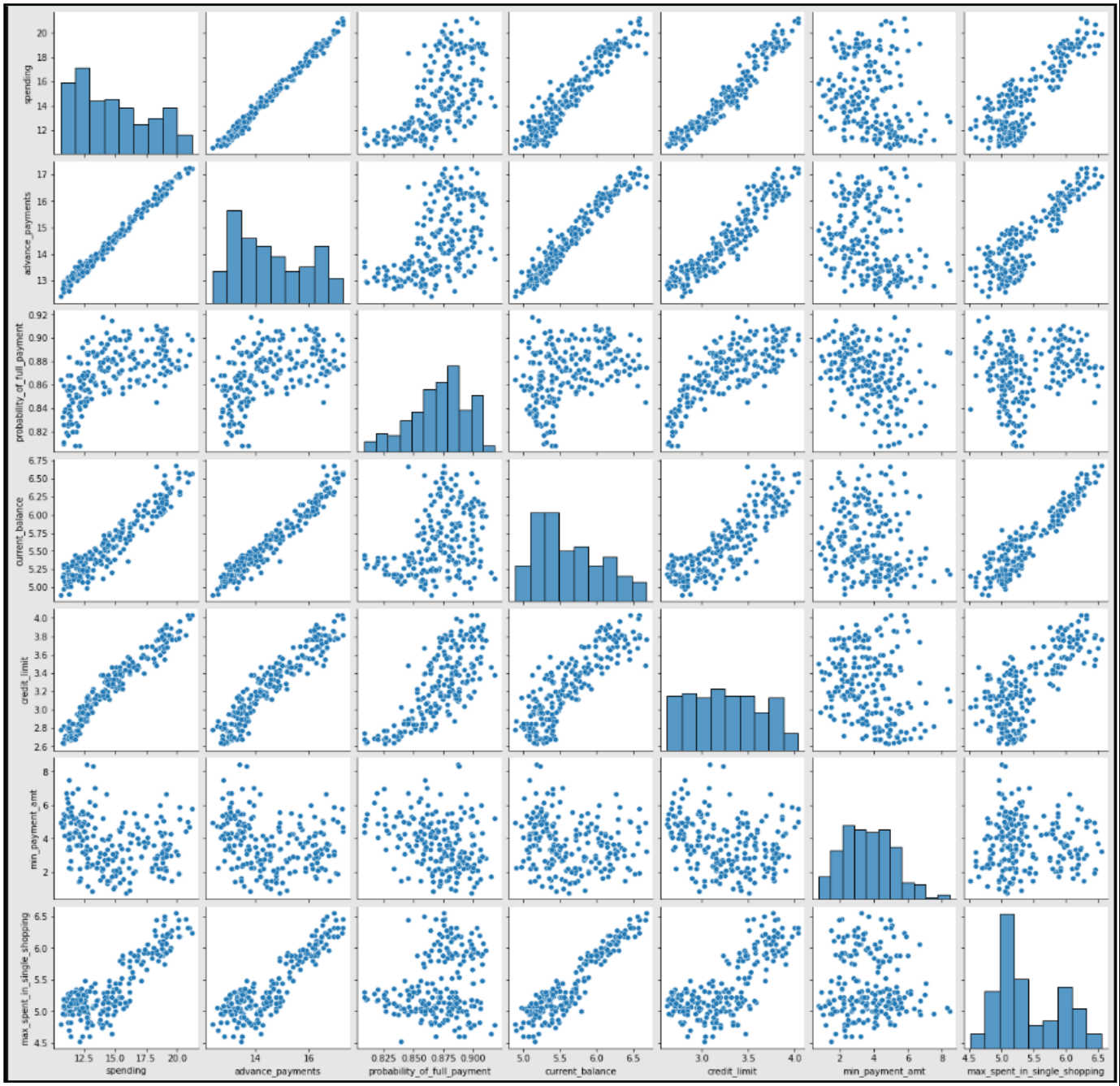

Figure 7: Scatter plot

Observation:

- There are mostly positive correlation is seen here among Spending Vs advance payments, Current balance.
- The spending of individual increases with their credit limit, current balance and advance payments. It can be further discussed in following heat map.

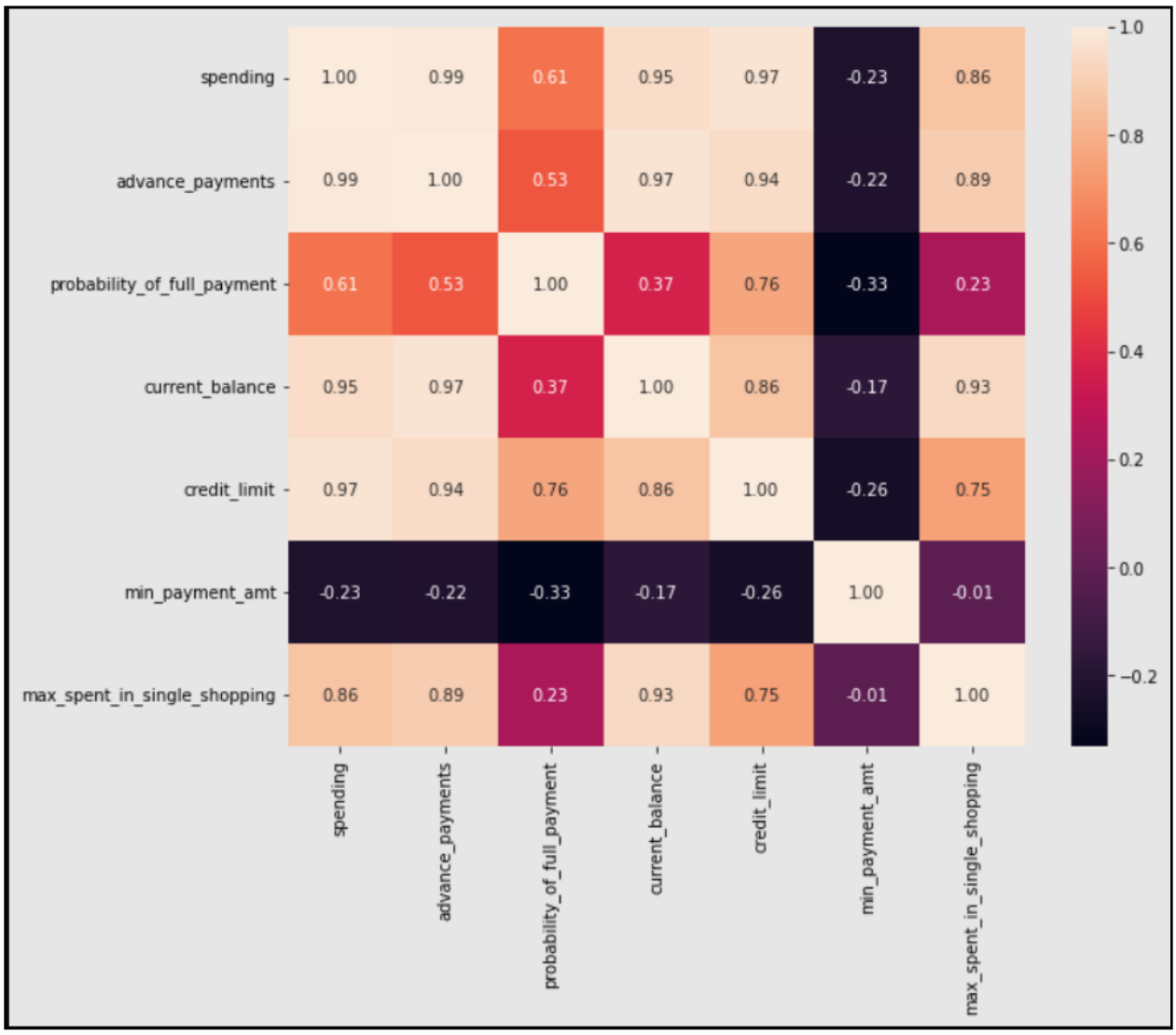

Figure 8: Heat map

Observation:

- The strong positive correlation is between Spending Vs advance payment with coefficient of 0.99(99%).
- The next stronger correlation is between Spending Vs credit limit with coefficient of 0.97(97%).
- The other stronger correlation is between Spending Vs current balance with coefficient of 0.95(95%).
- This shows that the spending amount of customer is directly proportional to the amount in their credit limit and current balance. As the spending

increases they continue to increase their current balance to compensate the expenses.

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

Solution:

Scaling is necessary for a dataset to change the values of numerical or categorical variable to follow a common scale.

Here Zscore scaling is used to make the dataset precisely range from +3 to -3 values.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02	2.100000e+02
mean	9.148766e-16	1.097006e-16	1.243978e-15	-1.089076e-16	-2.994298e-16	5.302637e-16	-1.935489e-15
std	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00	1.002389e+00
min	-1.466714e+00	-1.649686e+00	-2.668236e+00	-1.650501e+00	-1.668209e+00	-1.956769e+00	-1.813288e+00
25%	-8.879552e-01	-8.514330e-01	-5.980791e-01	-8.286816e-01	-8.349072e-01	-7.591477e-01	-7.404953e-01
50%	-1.696741e-01	-1.836639e-01	1.039927e-01	-2.376280e-01	-5.733534e-02	-6.746852e-02	-3.774588e-01
75%	8.465989e-01	8.870693e-01	7.116771e-01	7.945947e-01	8.044956e-01	7.123789e-01	9.563941e-01
max	2.181534e+00	2.065260e+00	2.006586e+00	2.367533e+00	2.055112e+00	3.170590e+00	2.328998e+00

Figure 9: Scaled data

Observation:

- Z-score scaling helps in standardizing the values in same scale and using this technique helps us understand the number of standard deviations above and below the mean that each value falls.
- Here the standard deviation is equal for all the variables.

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

Solution:

Data records are sequentially grouped to create clusters based on distances between records and distances between clusters. Hierarchical clustering produces a useful graphical display of the clustering process and results called a dendrogram.

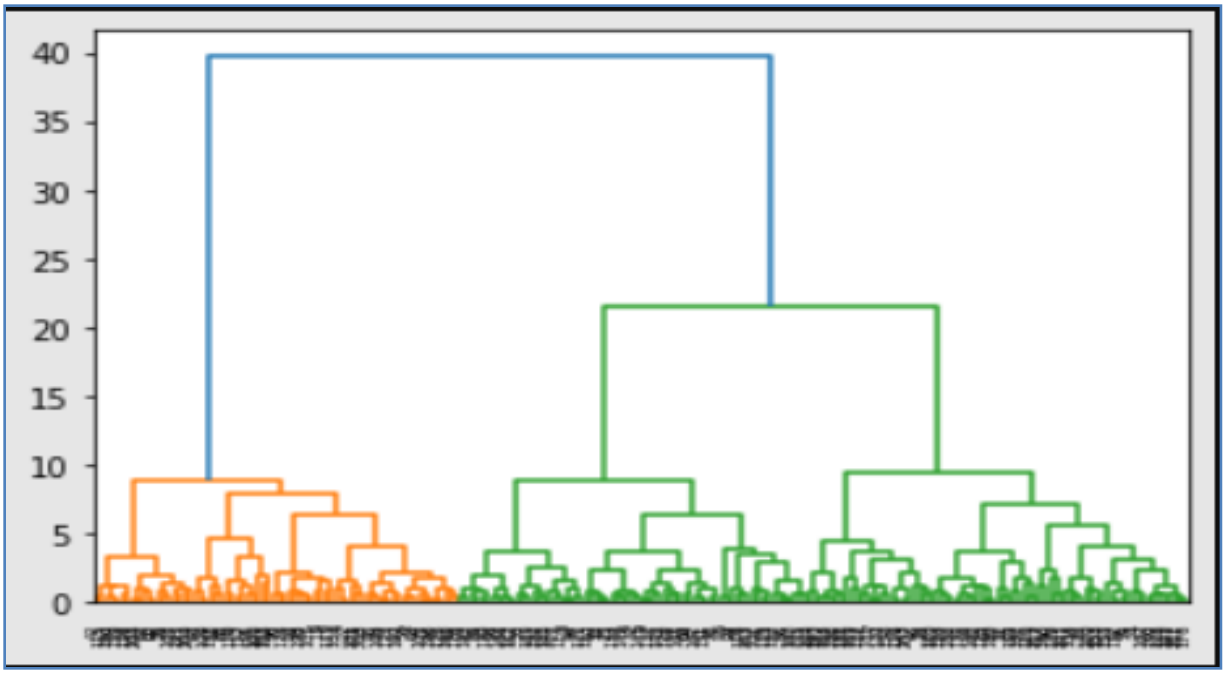

Figure 10: Dendrogram

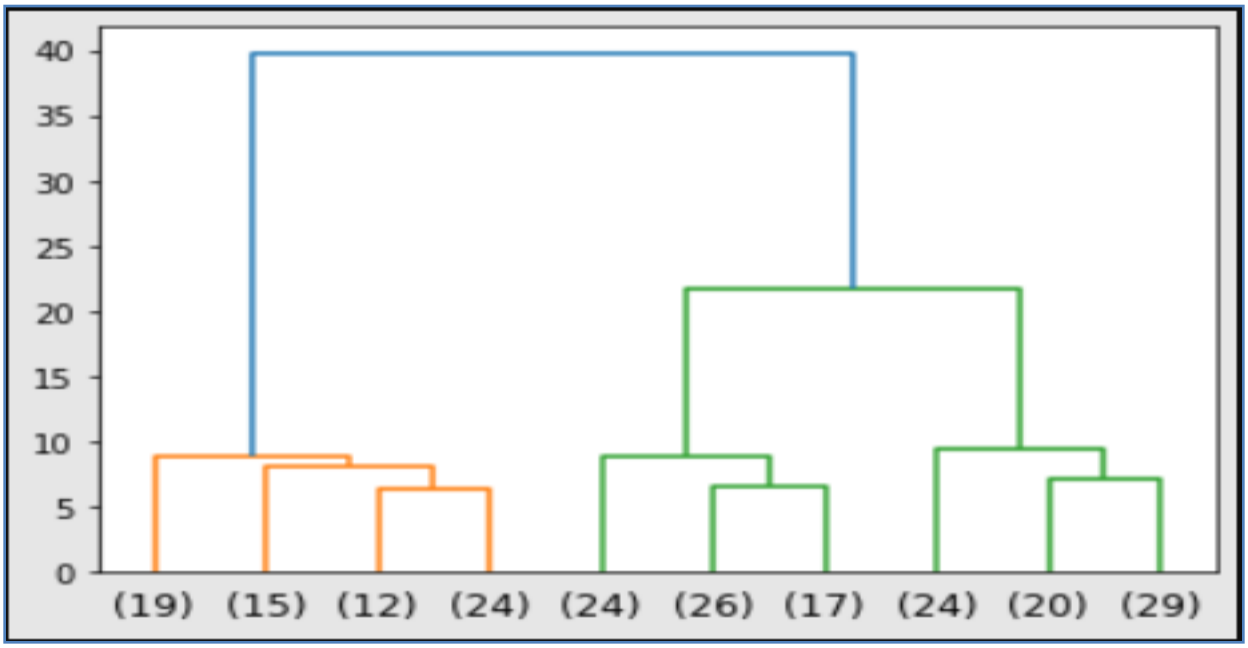

Figure 11:Truncated Dendrogram

- Here the truncated dendrogram shows the last 10 merged clusters of the data.
- Every horizontal line indicates the cluster that merges the number of observations under them with three different colours.

Cluster profile Observation:

- The optimum number of cluster is chosen as 3.
- The array of clusters shown below indicates each variable divided among three clusters one by one separated by comma.

Cluster-1:

Array([1, 3, 1, 2, 1, 2, 2, 3, 1, 2, 1, 3, 2, 1, 3, 2, 3, 2, 3, 2, 2, 2,
1, 2, 3, 1, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1,
2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 3, 3, 1,
1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 3, 3, 1,
1, 2, 3, 1, 3, 2, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 3, 1, 1, 2, 2, 1,
3, 3, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3,
3, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 2, 1, 2, 3, 2, 3, 2, 3, 3,
3, 3, 3, 2, 3, 1, 1, 2, 1, 1, 1, 2, 1, 3, 3, 3, 3, 2, 3, 1, 1, 1,
3, 3, 1, 2, 3, 3, 3, 3, 1, 1, 3, 3, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2,
1, 2, 3, 1, 3, 2, 1, 3, 1, 3, 1, 3] dtype=int32)

- **Ward's linkage** method helps to store the various distances of clusters that are sequentially merged into single large cluster.
- Maxclust criterion is used to group maximum number of clusters.
- Distance criterion - If a horizontal line is drawn referring to Y axis of dendrogram, it gives us the optimum number of clusters.
- Hierarchical clustering has the advantage of getting desired number of clusters by cutting the dendrogram at certain level.
- Both maxclust and distance criterion gives the optimum number of clusters.

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

Figure 12: H_Cluster

- The clusters attached to the dataset helps to identify the number of customers under the available categories.
- Cluster 1-**70 Customers**, Cluster 2- **67 Customers**, Cluster 3- **73 Customers** who are categorized as high/medium/low spending customers.

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.

Solution:

K-means clustering aims to partition data into number of clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart.

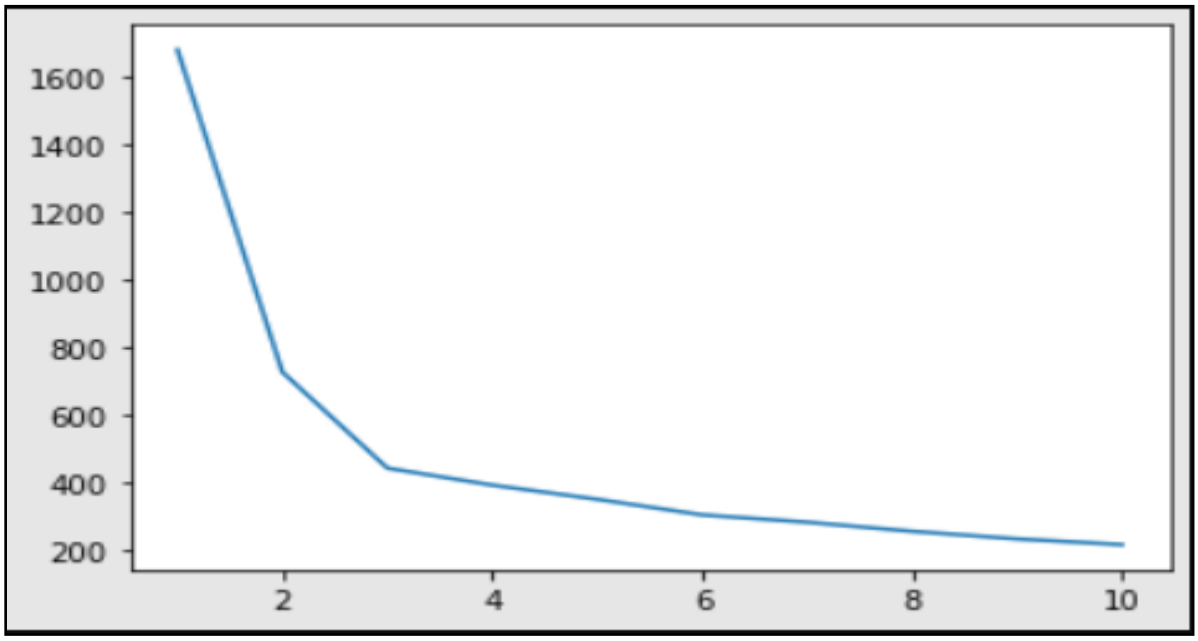

Figure 13: WSS plot

- WSS also known as distortion plot/error plot helps to know how many clusters are needed as output in K-means clustering.
- WSS score for first 10 observations are as follows.

  [1679.9999999999993, 726.6698839310626, 442.45237115842224
  392.2411164416579, 350.2653036722847, 304.1796519887737
  282.56820871436736, 255.60274170712367, 233.1807285338401
  216.24604850803541]

- The optimum number of cluster chosen here is 3 as there is a significant drop in WSS (Without sum of squares) value.

- Silhouette score is an indirect model evaluation technique that helps us to analyze if each observation to clusters are correctly connected are not using distance criterion.
- Sil_width= b-a/max(a,b) where b= distance between observation and the neighbour cluster centroid (C2),a= distance between observation and its own cluster centroid (C1).

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	H_clusters	clus_kmeans	sil_width
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1	1	0.611016
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3	2	0.458139
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1	1	0.689726
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2	0	0.536056
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1	1	0.519094

Figure 14:K_Means cluster

Inference:

The silhouette score is found to be positive (max=0.689,min=0.009) in the model which indicates the mapping of the observation is correct to its current centroid.

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

Solution:

H_clusters: 3 group clusters through hierarchical clustering.

H_clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178

Figure 15:H_cluster mean

Cluster 1: High spending group

Cluster 2: Low spending group

Cluster 3: Medium spending group

K_means Cluster: 3 group clusters through KMeans clustering.

clus_kmeans	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371
1	11.865882	13.255147	0.847857	5.238015	2.846632	4.909309	5.122353
2	14.237500	14.248889	0.879825	5.482431	3.233500	2.617614	5.085542

Figure 16:KMeans mean

Cluster 0: High spending group

Cluster 1: Low spending group

Cluster 2: Medium spending group

Recommendations:

- Low spending groups - customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.
- Medium spending groups -They are potential customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate to them.
- High Spending groups- These customers are very timely in their payments and giving any reward points might increase their purchases. Additional credit limit and loan offers will increase the attraction of more customers to the bank.

Problem Statement 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.

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).

Solution:

Shape: (3000, 10)

Data type: Int,Float,Object

Length: **10** Columns

Missing values: No missing values in the dataset.

Summary dataset: Count, mean, std.deviation, range of values and IQR range.

The dataset has 10 columns and 3000 entries with no missing data and all the entries are filled with int,float and object data types.

	Age	Commision	Duration	Sales
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	14.529203	70.001333	60.249913
std	10.463518	25.481455	134.053313	70.733954
min	8.000000	0.000000	-1.000000	0.000000
25%	32.000000	0.000000	11.000000	20.000000
50%	36.000000	4.630000	26.500000	33.000000
75%	42.000000	17.235000	63.000000	69.000000
max	84.000000	210.210000	4580.000000	539.000000

Figure 17: Summary dataset

Observation:

- Based on the above figure, the data looks good with no missing variables.

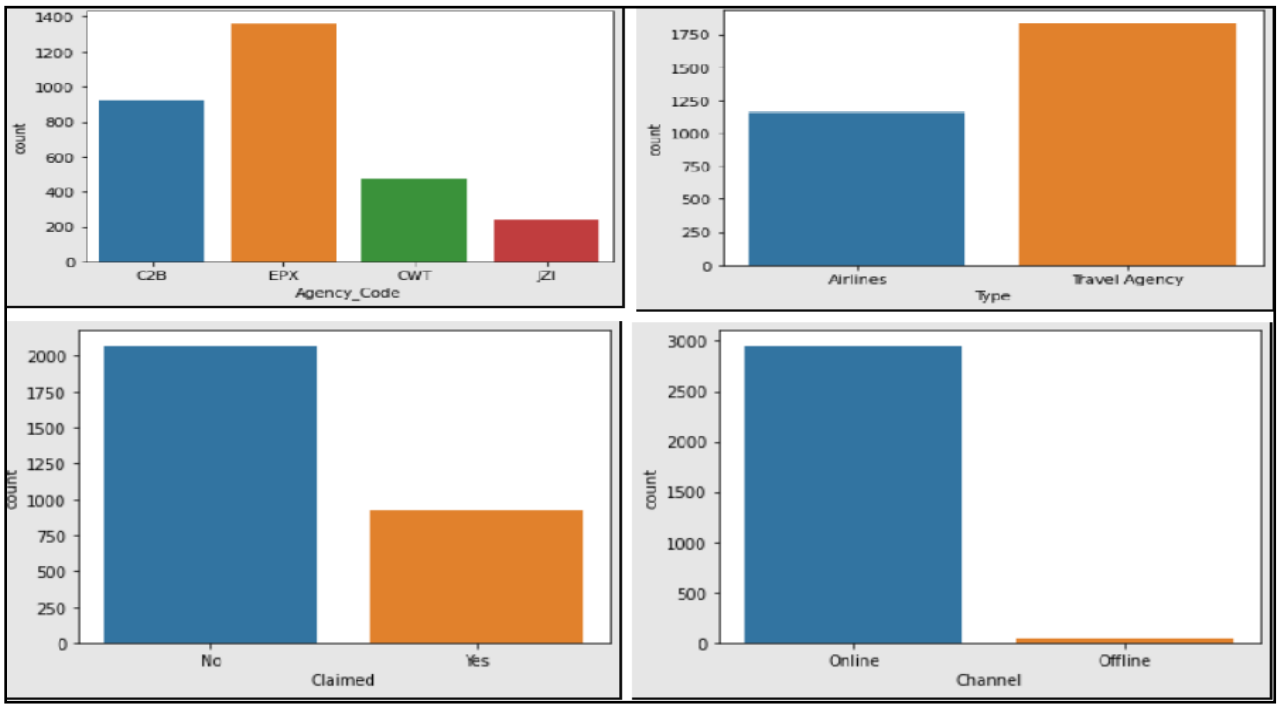

Figure 18:Countplot

Observation:

- Countplot for categorical variables gives us insight of the attributes with maximum and minimum values present in dataset.
- Agency_code-Customers under EPX code are higher than the rest codes.
- Type- Most customers prefer travel agency than airlines.
- Claimed-Majority of customers have opted no to travel insurance.
- Channel- Most travel distributions are made through online.

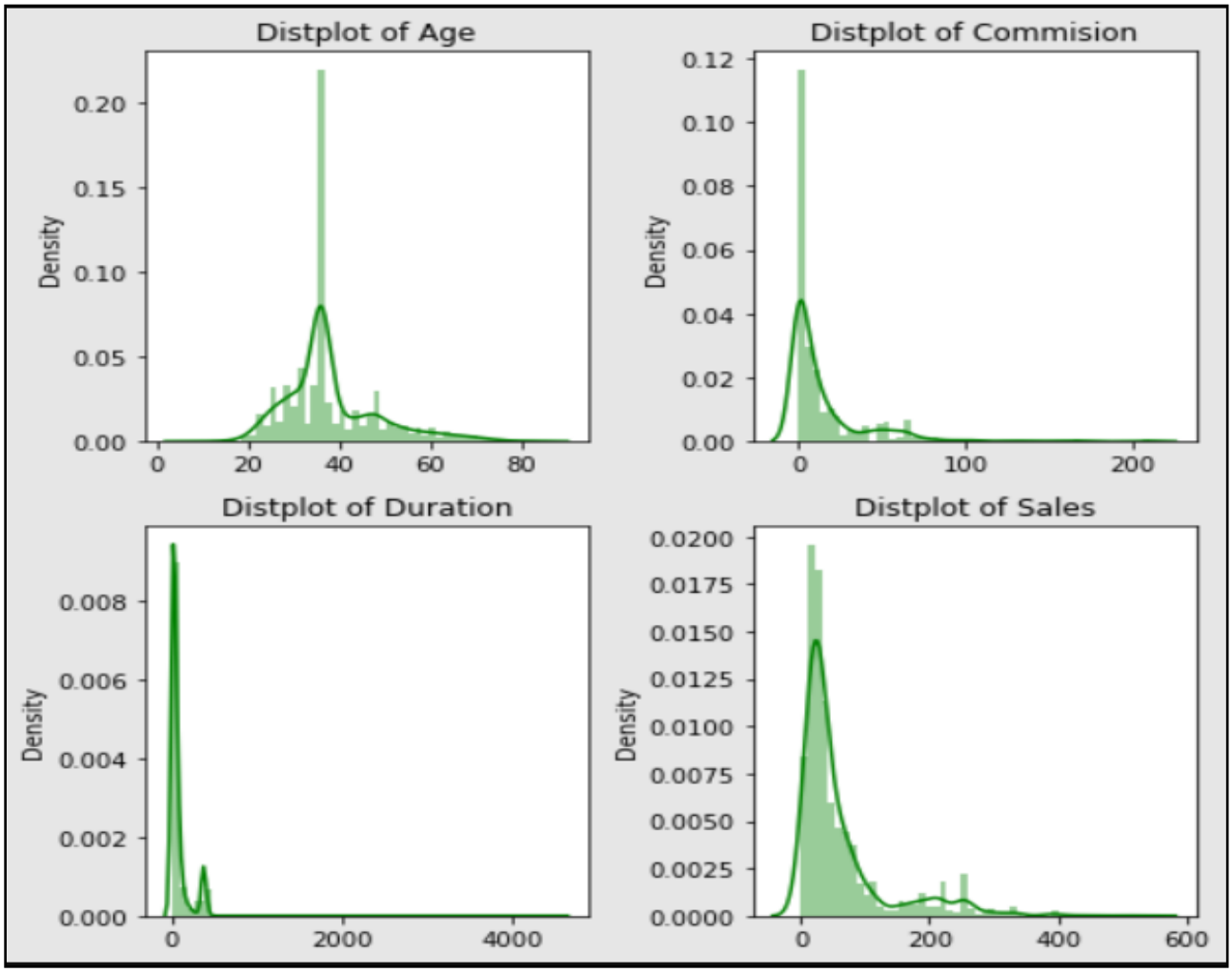

Figure 19: Distplot

Observation:

- The Distplot distribution is almost right skewed for all the variables.

- People with age group of 30-45 tends to use travel agency more than other age groups.
- Commission received for insurance firm is in less percentage.
- Duration of stay only lasts for days.
- Amount of sales is also less than 200 as shown in distplot.

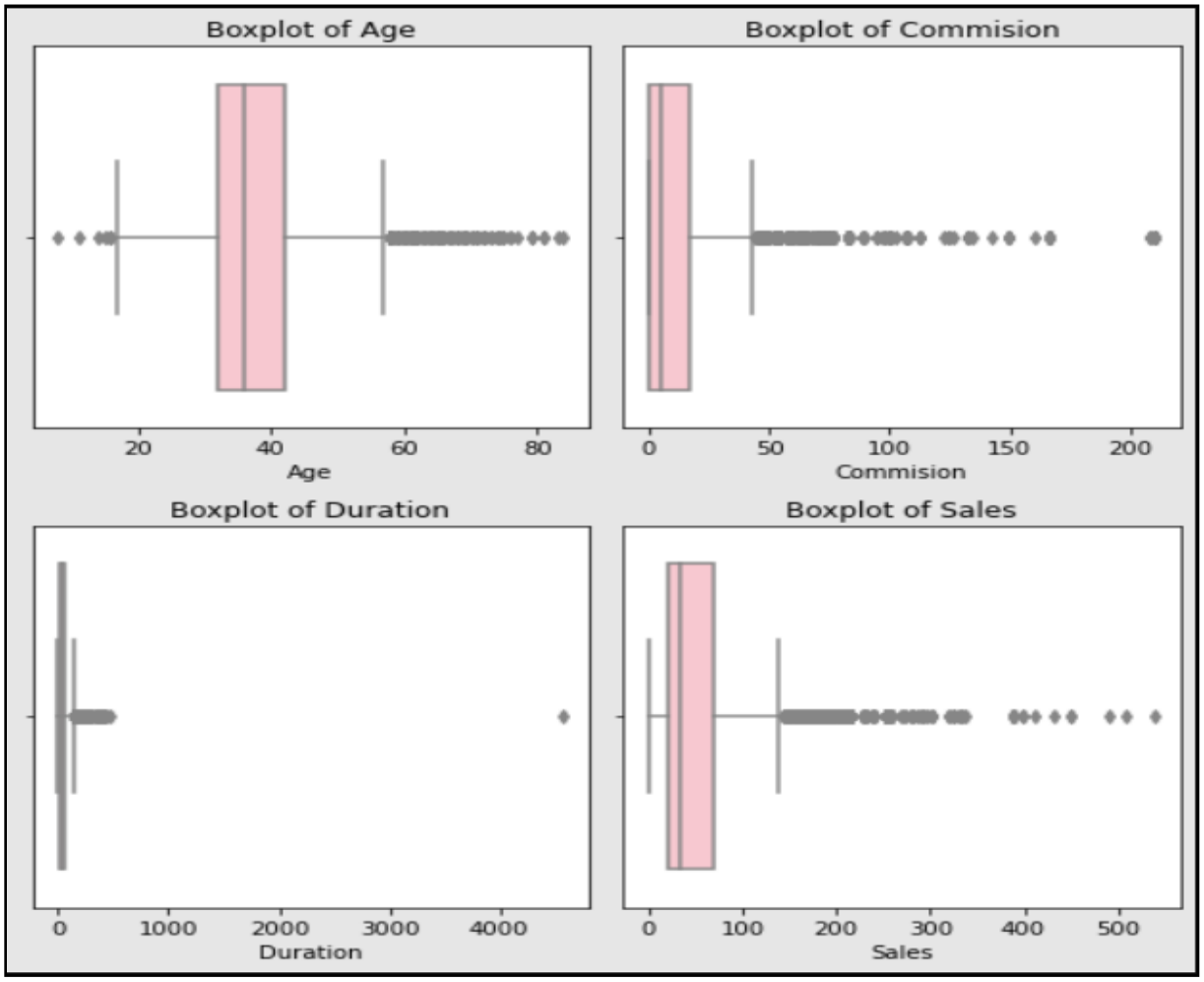

Figure 20: Boxplot

Observation:

- Age: Lower outlier 17 and Upper outlier 27
- Commission: Lower outlier -25.85 and Upper outlier 43.08
- Duration: Lower outlier -67 and Upper outlier 141
- Sales: Lower outlier -53.5 and Upper outlier 142.5
- The outliers present are retained in the dataset.

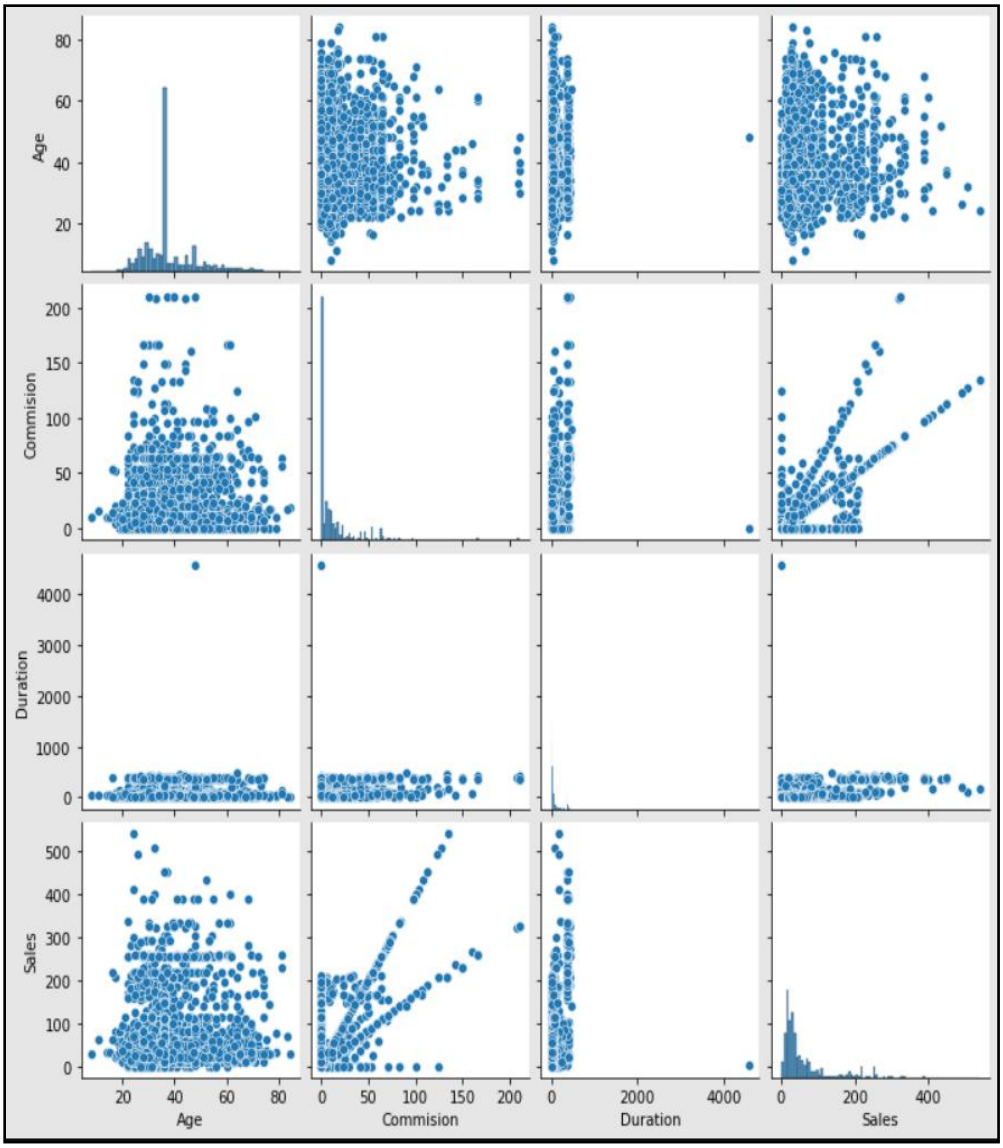

Figure 21:Pairplot

Observation:

- Positive correlation is seen between Commission Vs Sales as sales increases commission to the travel firm also increases.
- Negative correlation is with duration of stay as people prefer not to stay longer in their destination. It is very fewer compared to other attributes.

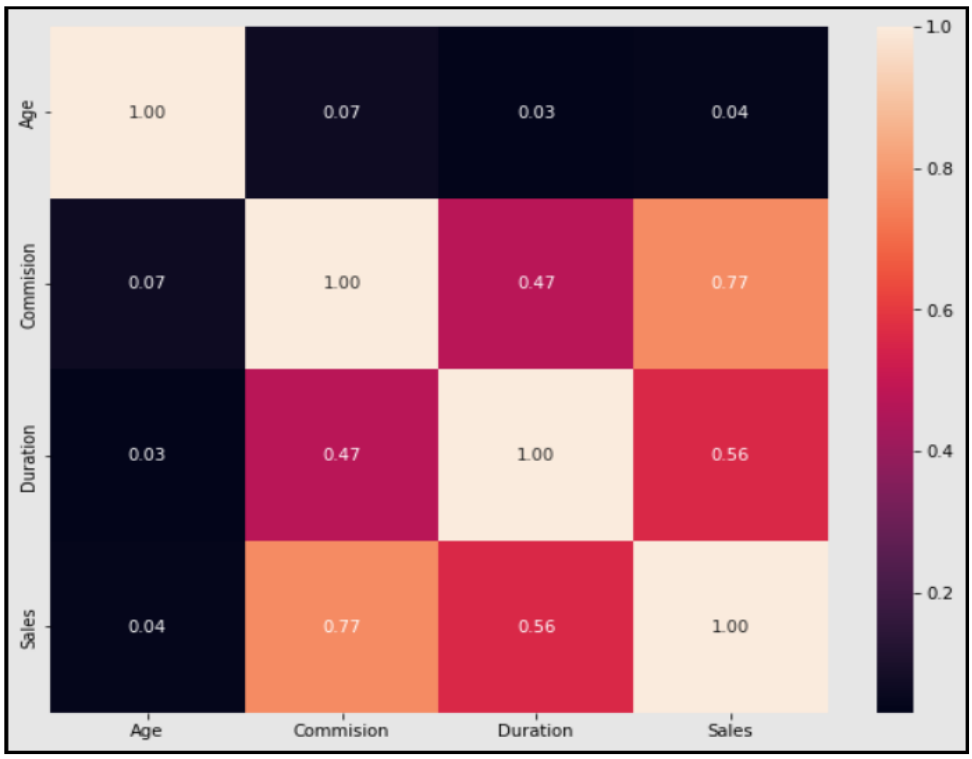

Figure 22: Correlation plot

Observation:

- The strong positive correlation is between Commission Vs Sales with coefficient of 0.77(77%).
- The next stronger correlation is between Duration Vs Sales with coefficient of 0.56(56%).
- The other stronger correlation is between Duration Vs Commission with coefficient of 0.47(47%).

The above correlation shows that Sales is directly proportional to commission and duration of the stay in travel.

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

Solution:

CART- A decision Tree is one of most popular and effective supervised learning technique for classification problem that equally works well with both categorical and quantitative variables. It is a graphical representation of all the possible

solution to a decision that is based on certain condition. In this algorithm, the training sample points are split into two or more sets based on the split condition over input variables.This method is very easy to interpret and no data preprocessing is required.

This is a Binary Decision Tree and splitting criteria used here is "Gini Index"

Train and Test data shapes: X_train-(2100,9) & X_test-(900,9)

CART-Tree:

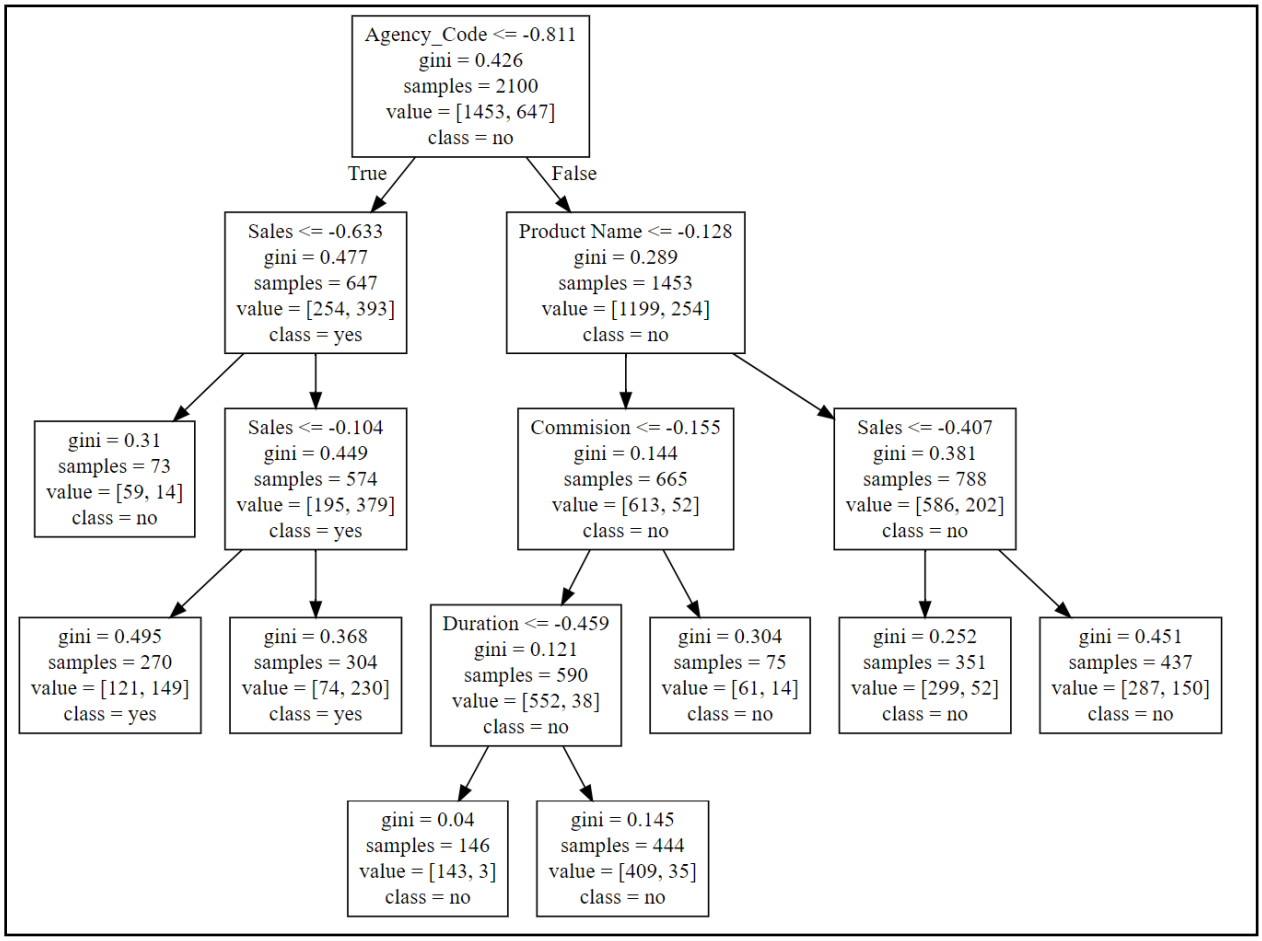

Figure 23:CART-Tree

From the above tree,the gini index is found to be lower which indicates that there is lesser impurity in the model.

Feature Importance of CART:

The prediction of test data can be found using “Feature importance” function. The following data with zeros can be eliminated and other values are taken into consideration.

	Imp
Agency_Code	0.674494
Sales	0.222345
Product Name	0.092149
Commision	0.008008
Duration	0.003005
Age	0.000000
Type	0.000000
Channel	0.000000
Destination	0.000000

Figure 24:CART-Imp

Train and Test data Prediction:

The good and bad data in terms of 0’s and 1’s can be predicted using probability function as shown below.

	0	1
0	0.656751	0.343249
1	0.979452	0.020548
2	0.921171	0.078829
3	0.656751	0.343249
4	0.921171	0.078829

Figure 25: Prob-CART

Random Forest- This is an ensemble technique wherein we construct multiple models and take the average output of all the models to take final decision.

	0	1
0	0.777642	0.222358
1	0.980005	0.019995
2	0.896432	0.103568
3	0.647552	0.352448
4	0.887602	0.112398

Figure 26: Prob-RFCL

	Imp
Agency_Code	0.331032
Product Name	0.190926
Sales	0.181657
Commision	0.113425
Type	0.070799
Duration	0.063841
Age	0.038955
Destination	0.007602
Channel	0.001762

Figure 27: Imp-RFCL

ANN - A Machine Learning algorithm that is roughly modelled around what is currently known about how the human brain functions.

Artificial Neural Network models the relationship between a set of input signals and an output. It uses a network of artificial neurons or nodes to solve challenging learning problems.

	0	1
0	0.822676	0.177324
1	0.933407	0.066593
2	0.918772	0.081228
3	0.688933	0.311067
4	0.913425	0.086575

Figure 28:Prob-ANN

Observation:

- Based on the above models, the data fits good with Random Forest Model with no bad data in it.

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.

Solution:

CART-Train and Test Data:

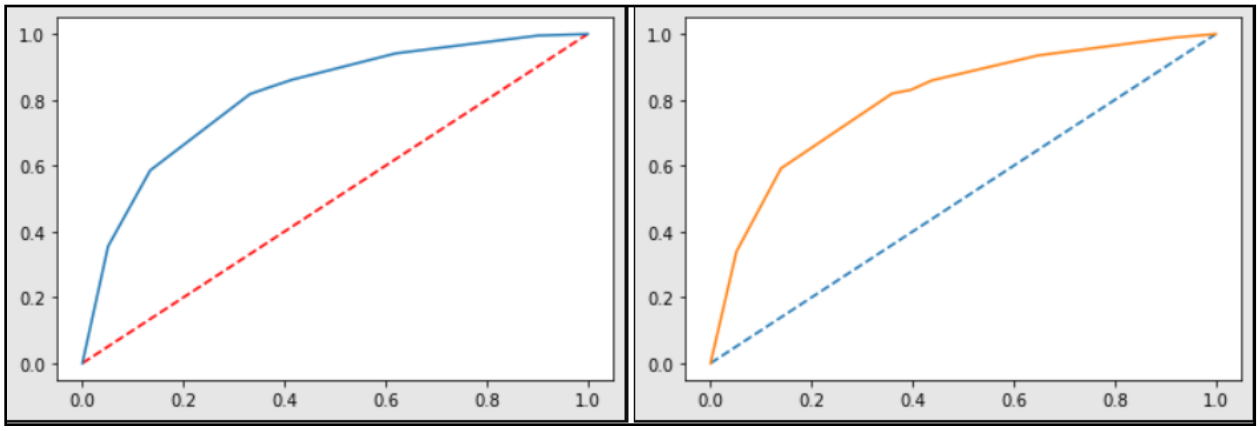

Figure 29:CART-ROC Curves

- AUC score for Train and Test: 0.81 & 0.80

- CART-Confusion Matrix(Train): Array[1258,195

  268,379]

- ACC Score(Train):0.77

	precision	recall	f1-score	support
0	0.82	0.87	0.84	1453
1	0.66	0.59	0.62	647
accuracy			0.78	2100
macro avg	0.74	0.73	0.73	2100
weighted avg	0.77	0.78	0.78	2100

Figure 30: CART Train Metrics

- CART-Confusion Matrix(Test): Array[536,87

  113,164]

- ACC Score(Test):0.77

	precision	recall	f1-score	support
0	0.83	0.86	0.84	623
1	0.65	0.59	0.62	277
accuracy			0.78	900
macro avg	0.74	0.73	0.73	900
weighted avg	0.77	0.78	0.77	900

Figure 31: CART-Test Metrics

RF-Train and Test Data:

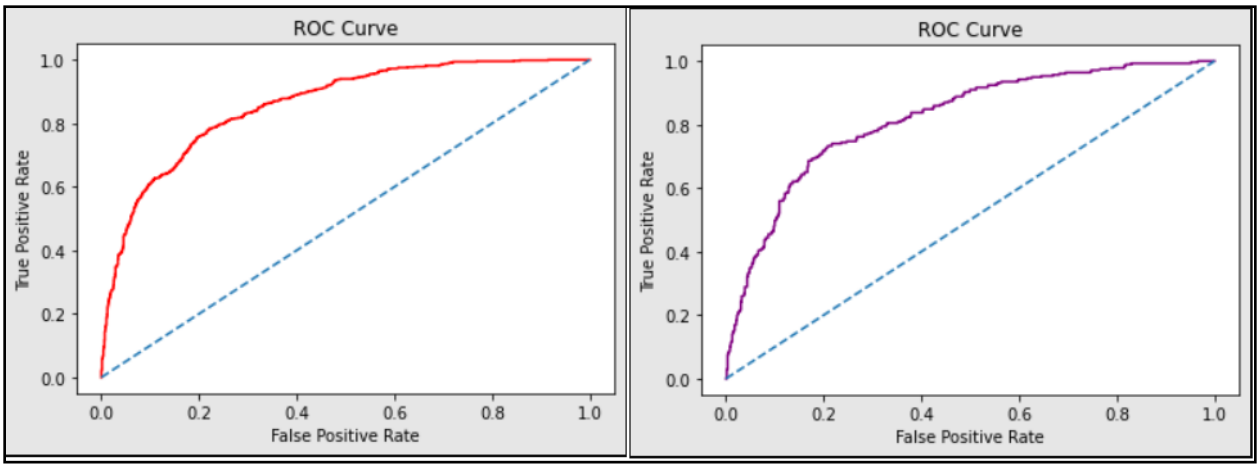

Figure 32: RF-ROC Curves

- AUC score for Train and Test: 0.85 & 0.81
- RF-Confusion Matrix(Train): Array[1291,162

  241,406]

- ACC Score(Train):0.80

	precision	recall	f1-score	support
0	0.84	0.89	0.86	1453
1	0.71	0.63	0.67	647
accuracy			0.81	2100
macro avg	0.78	0.76	0.77	2100
weighted avg	0.80	0.81	0.80	2100

Figure 33:RF-Train Metrics

- RF-Confusion Matrix(Test): Array[547,76

  115,162]

- ACC Score(Test):0.78

	precision	recall	f1-score	support
0	0.83	0.88	0.85	623
1	0.68	0.58	0.63	277
accuracy			0.79	900
macro avg	0.75	0.73	0.74	900
weighted avg	0.78	0.79	0.78	900

Figure 34:RF-Test Metrics

ANN-Train and Test Data:

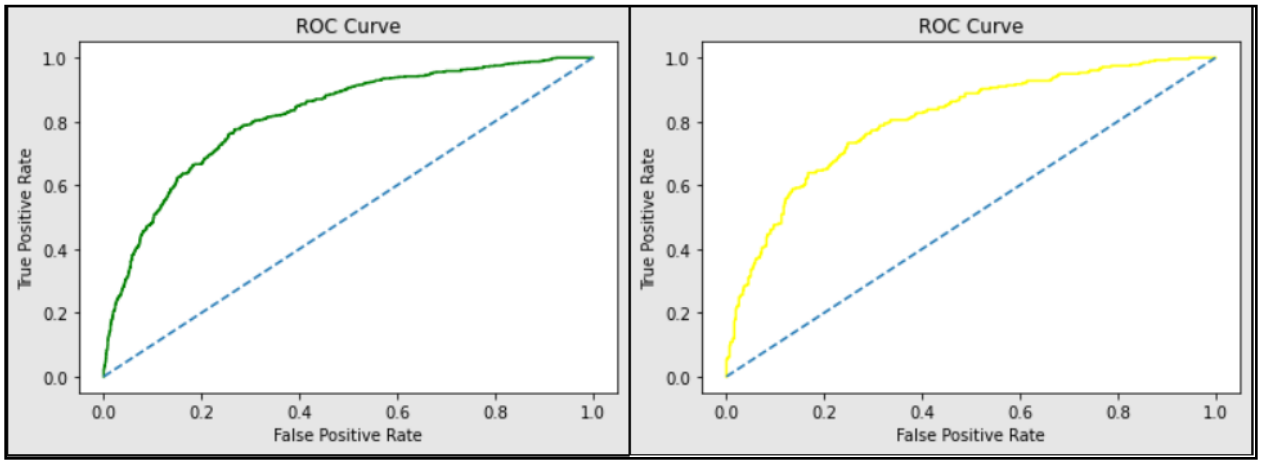

Figure 35: ANN-ROC Curves

- AUC score for Train and Test: 0.81 & 0.80
- ANN-Confusion Matrix(Train): Array[1298,155

  315,332]

- ACC Score(Train):0.77

	precision	recall	f1-score	support
0	0.80	0.89	0.85	1453
1	0.68	0.51	0.59	647
accuracy			0.78	2100
macro avg	0.74	0.70	0.72	2100
weighted avg	0.77	0.78	0.77	2100

Figure 36:ANN-Train Metrics

- ANN-Confusion Matrix(Test): Array[553,70

  138,139]

- ACC Score(Test):0.76

	precision	recall	f1-score	support
0	0.80	0.89	0.84	623
1	0.67	0.50	0.57	277
accuracy			0.77	900
macro avg	0.73	0.69	0.71	900
weighted avg	0.76	0.77	0.76	900

Figure 37:ANN-Test Metrics

Observation:

- The train and test data for all the three models are almost similar to each other. Hence the models are identified as a good fit.
- Precision and Recall are positive among all the three models
- From AUC & ROC for training data-the probability under “Positive” class is stronger and evident.
- Upon comparing all data there is no overfitting found in these models.

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

Solution:

	C-Train	C-Test	RF-Train	RF-Test	NN-Train	NN-Test
Accuracy	0.77	0.77	0.80	0.78	0.77	0.76
AUC	0.81	0.80	0.85	0.81	0.81	0.80
Precision	0.66	0.65	0.71	0.68	0.68	0.67
Recall	0.59	0.59	0.63	0.58	0.51	0.50
F1 Score	0.62	0.62	0.67	0.63	0.59	0.57

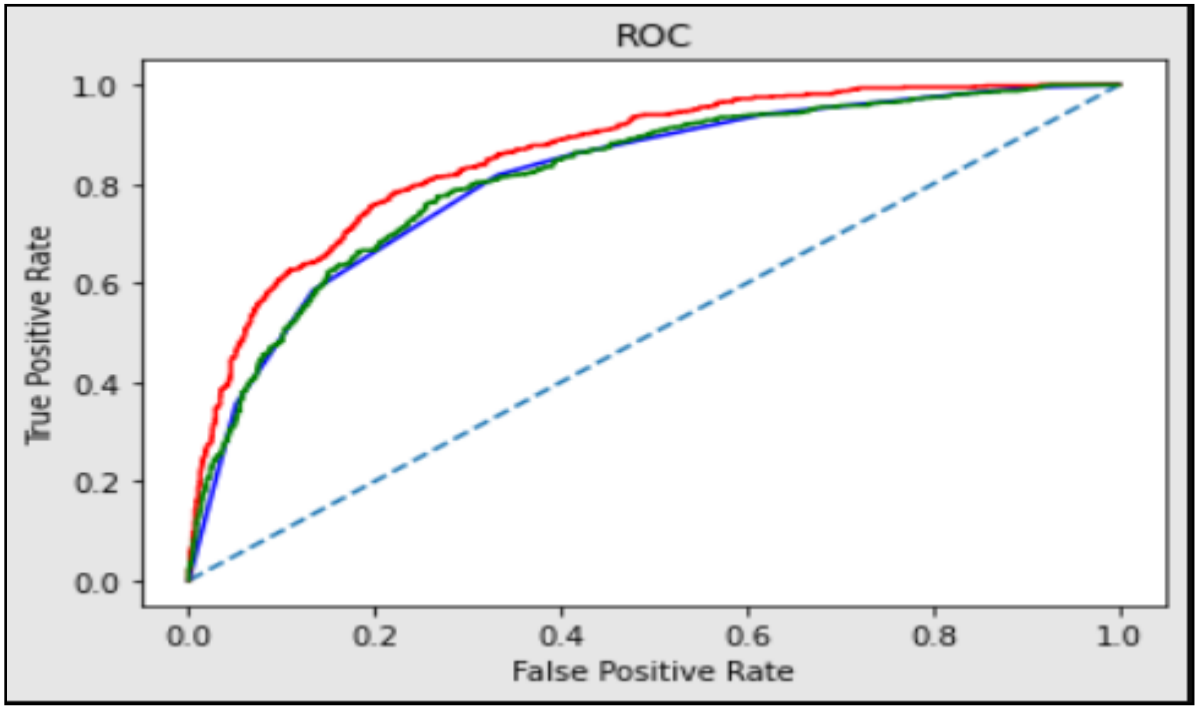

Figure 38:ROC-Train

Labels:

Blue line- CART model, Red line- RF model, Green line-ANN Model.

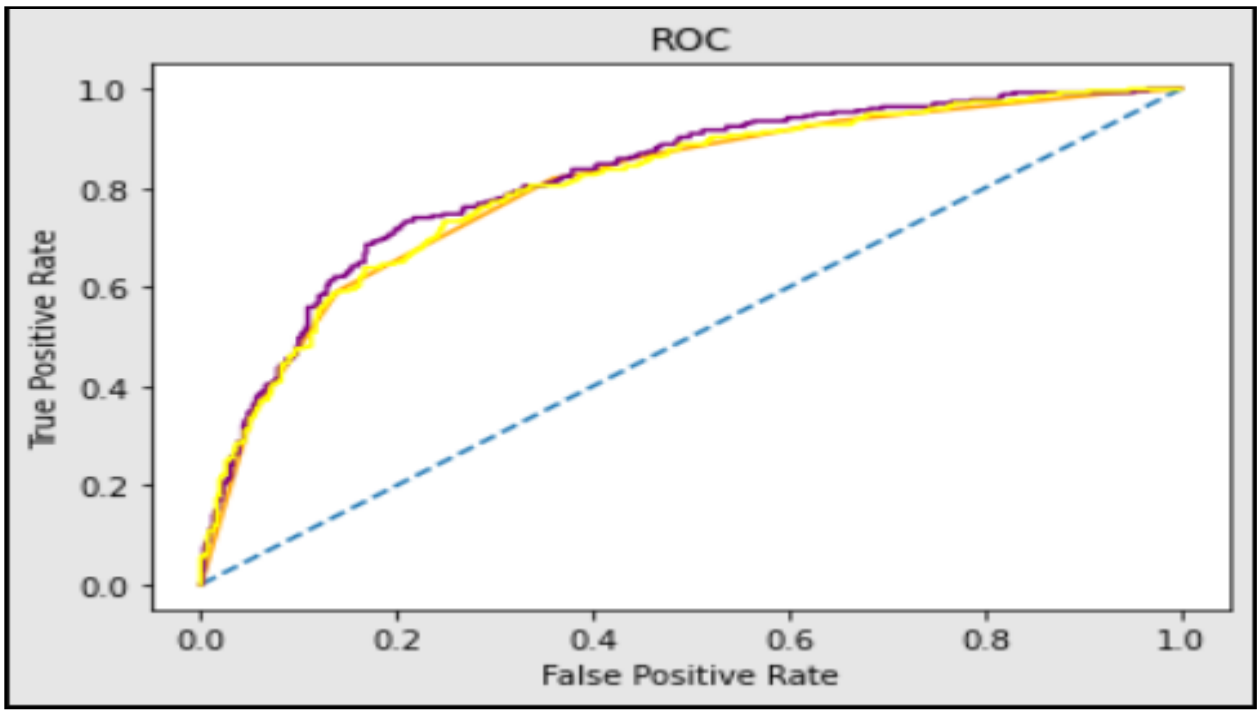

Figure 39: ROC Test

Labels:

Orange line- CART model, Purple line- RF model, Yellow line-ANN Model.

Observation:

- The train and test ROC curves for all the three models are combined.
- Out of these RF model is best as its F1 score is better than other models.

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

Solution:

- From the plots 90% of insurance is done by online channel and the claimed people are from offline category.
- The JZI agency should improve sales as they are less compared to other agencies.
- Frequent promotions and advertisements will help increase the sales.

- The destination to America and Europe will increase the sale and commission to travel agencies by providing attractive offers to choose these destinations.
- Since the accuracy of the dataset is around 80%,the overall increase in travel through Airlines will draw more spending and relevant insurance can be increased.

**********Thank You**********