Problem_ 1: Clustering

Table of Contents

Contents

Executive Summary

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of clustering techniques convert data are group wise customer past month credit card use .Past month credit card spending summarize data bank received payment and min limit payment amount , max spent in single shopping 1000s and min are 100s. Using hierarchical clustering and K-men for find group of customer performance.

Data Description

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)	

Sample of the dataset

	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

Fig_0.1

Conclusion | insight: This is Sample of given data all information is shown in this manner

Exploratory Data Analysis

```
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
```

Fig_0.2

Conclusion | insight: Seven type of information in this data seat spending, advance payments probability_of_full_payment ,current_balance ,credit limit and min_payment_amt ,max_spent_in_single_shopping.data is 11.6 kb and 0 to 210 row and 7 columns.

Check for missing values in the dataset

```
spending                      0
advance_payments              0
probability_of_full_payment   0
current_balance               0
credit_limit                  0
min_payment_amt               0
max_spent_in_single_shopping  0
dtype: int64
```

Fig_0.3

Conclusion | insight: From the above results we can see that there is no missing value present in the dataset.

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

Solution-

Univariate analysis

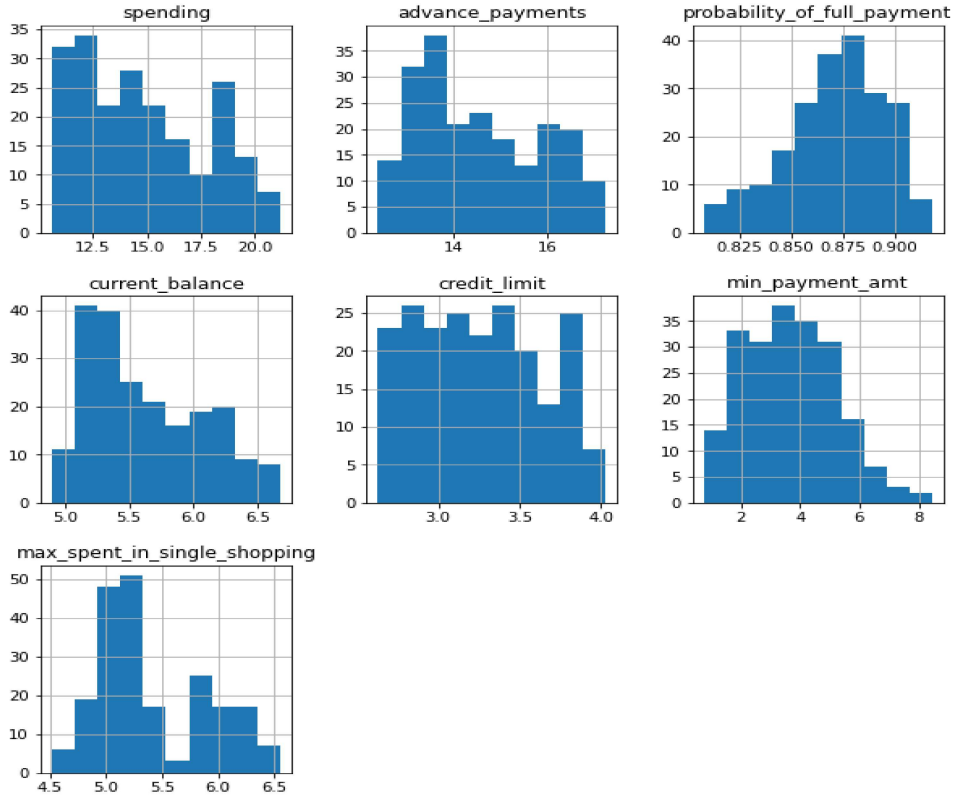

Fig_0.4

Fig_0.5

Conclusion | insight: All columns data single behaviour show in Fig_0.4 & Fig_0.5.

Bivariate Analysis

Column- spending graph

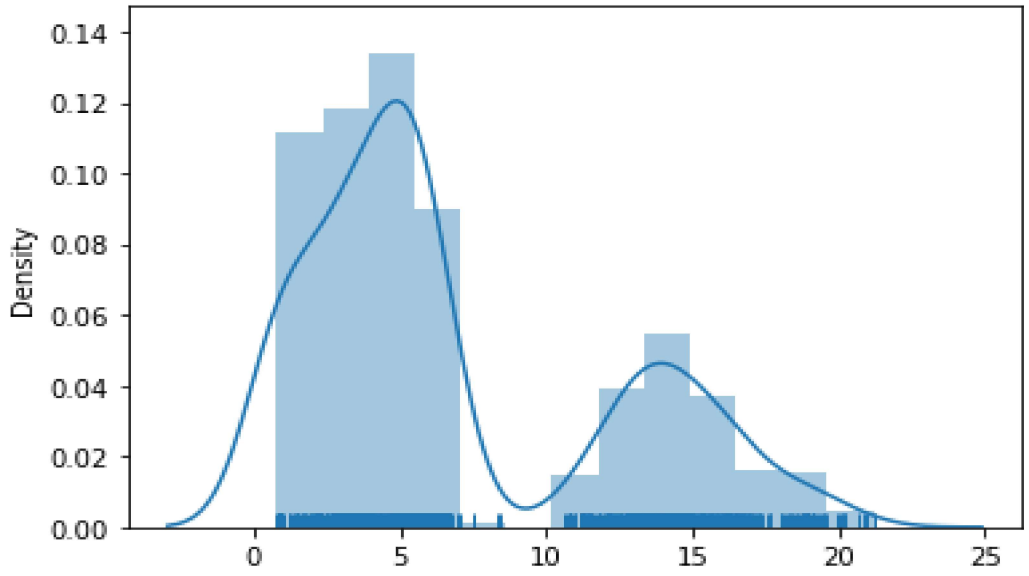

Fig_0.6

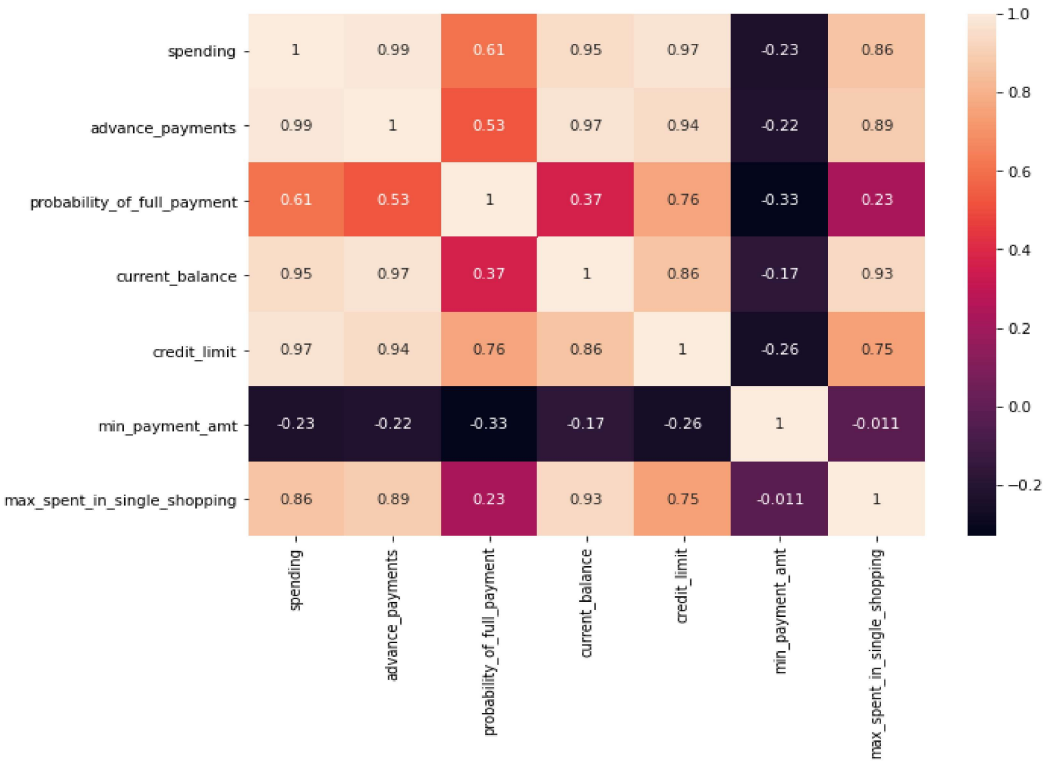

Fig_0.7

Conclusion | insight: Correlation of all two data behaviour in show in fig_0.7.

Multivariate Analysis

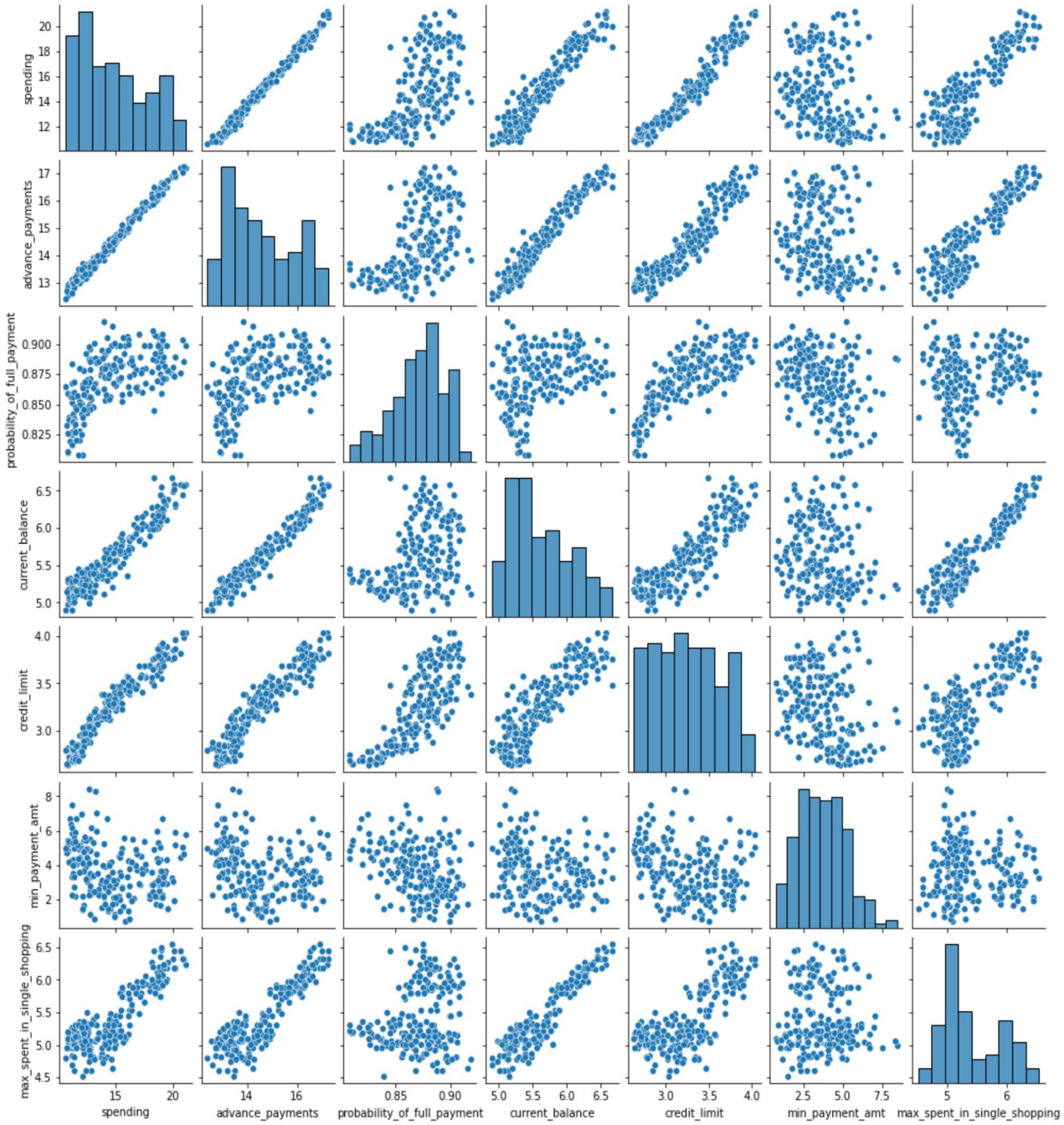

Fig_0.8

Conclusion | insight: All data pair correction and behaviour show in fig_0.8.

Q1.2 Do you think scaling is necessary for clustering in this case? Justify.

Solution-

Before scale Describe data

	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

Fig_0.8.1

After scale Describe data

	0	1	2	3	4	5	6
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	1.260896e-15	-1.358702e-16	-2.790757e-16	5.418946e-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

Fig_0.9

Conclusion | insight: yes, After looking data scaling is necessary for clustering for k-Mean and Hierarchical clustering also, As per describe data max/Min value Spending 21.18/10.59, and min_payment_ant 8.45/0.765 are magnitude all data are not similar, so scaling is required .And other measurement was Stander deviation compare for all data Std was variation to 0.02 to 2.90 in all group data was differ, so scaling is required. Scaling is mostly required for data calcite distance between two and more data. Working of scaling techniques for treat data for all distance the range of all features should be normalized after contributes approximately proportionately to find distance is that region after doing clustering for best result. . So we can go for scaling to avoid any performance issue due to the unscaled dataset.

Q1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Solution-

After Scaled data model

	0	1	2	3	4	5	6
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813
...	...	...	...	...	...	...	...
205	-0.329866	-0.413929	0.721222	-0.428801	-0.158181	0.190536	-1.366631
206	0.662292	0.814152	-0.305372	0.675253	0.476084	0.813214	0.789153
207	-0.281636	-0.306472	0.364883	-0.431064	-0.152873	-1.322158	-0.830235
208	0.438367	0.338271	1.230277	0.182048	0.600814	-0.953484	0.071238
209	0.248893	0.453403	-0.776248	0.659416	-0.073258	-0.706813	0.960473

Fig_10

After Scaling, magnitude of all the data is same and not varying very much as earlier.

Dendrogram of ward method all pointer graph

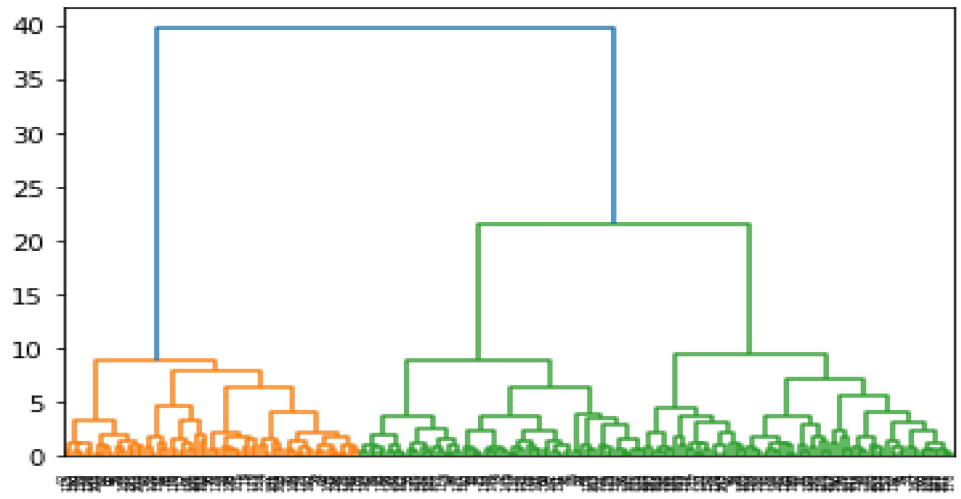

Fig_11

Dendrogram of ward method using 10 pointer graph

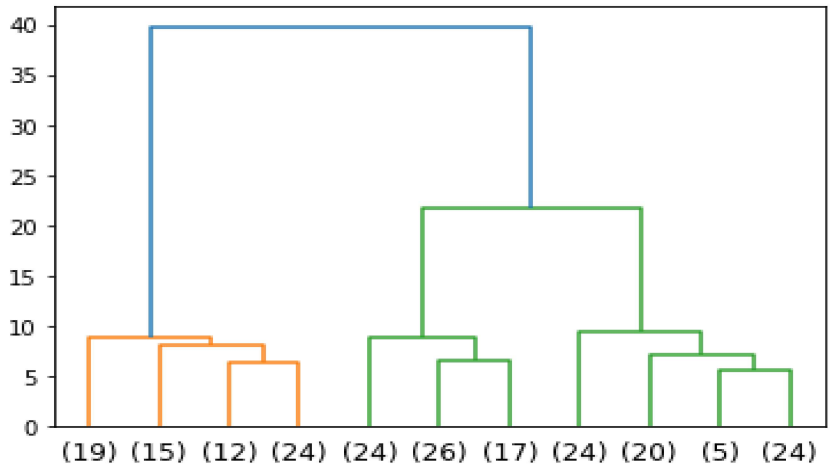

Fig_11

I have Used Ward linkage method to link the features of the dataset. This is truncated Dendrogram with P= 10. Since the original dendrogram was not clear at x-axis, so it was advisable to truncate it and get a neater diagram. We can also calculate the number of observations in each cluster by summing the numbers at x-axis. From dendrogram it is clear that no of optimum clusters should be 2.

Hierarchical clustering value count of clusters

```
1    70
2    67
3    73
Name: clusters, dtype: int64
```

Fig_12

Doing Cluster profiling for the 3 clusters as obtained from Hierarchical clustering clustering and taking their mean and generating the data frame for profiles.

Final result add clusters in main data

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	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

Fig_13

Cluster 1: Customer with high spendings, and having less probability of full payment.

Cluster 2: Customer with Medium spendings and having high probability of full payment

Cluster 3: Customer with low spendings and having high probability of full payment

Conclusion | insight: Based on the dataset, 3 clusters can be optimum to do customer segmentation as it is also viable from the dendrogram in Full Payer and Revolvers. Because, Non payers are not available in the dataset.

Q1.4 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-

KMeans clusters three labels

```
[2, 0, 2, 1, 2, 1, 1, 0, 2, 1, 2, 0, 1, 2, 0, 1, 0, 1, 1, 1, 1, 1,
 2, 1, 0, 2, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 2, 2, 0, 2, 2,
 1, 1, 0, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 2, 0, 1, 1, 0, 0, 2,
 2, 0, 2, 1, 0, 1, 2, 2, 1, 2, 0, 1, 2, 0, 0, 0, 0, 2, 1, 0, 2, 0,
 2, 1, 0, 2, 0, 1, 1, 2, 2, 2, 1, 2, 0, 2, 0, 2, 0, 2, 2, 1, 1, 2,
 0, 0, 2, 1, 1, 2, 0, 0, 1, 2, 0, 1, 1, 1, 0, 0, 2, 1, 0, 0, 1, 0,
 0, 2, 1, 2, 2, 1, 2, 0, 0, 0, 1, 1, 0, 1, 2, 1, 0, 1, 0, 1, 0, 0,
 1, 0, 0, 1, 0, 2, 2, 1, 2, 2, 2, 1, 0, 0, 0, 1, 0, 1, 0, 2, 2, 2,
 0, 1, 0, 1, 0, 0, 0, 0, 2, 2, 1, 0, 0, 1, 1, 0, 1, 2, 0, 2, 2, 1,
 2, 1, 0, 2, 0, 1, 2, 0, 2, 0, 0, 0])
```

Fig_12

Magnitude of all the data is on same scale. There is not a large difference between the magnitude of higher and lower values which essentially is essence of scaling.

Using K - Elbow Method

	0
0	1470.000000
1	659.171754
2	430.658973
3	371.301721
4	326.367602
5	289.420125
6	263.699963
7	241.319111
8	223.121257
9	205.184198

Fig_13

WSS Range graph as per K - Elbow Method

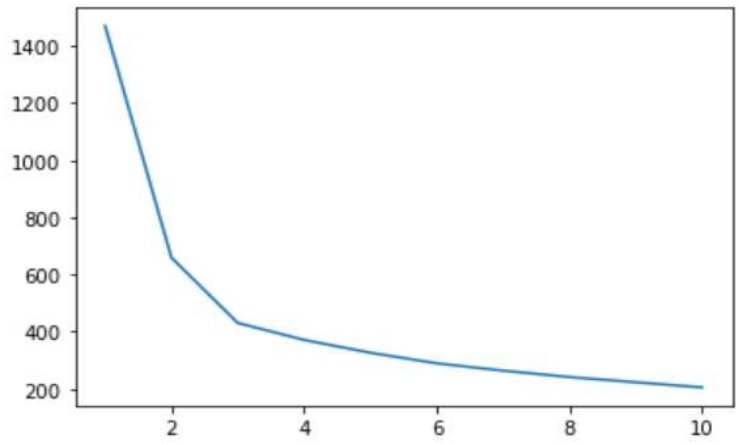

Fig_14

Let's go for Silhouette score to verify the optimum no of clusters, if the silhouette score for k=2 i.e. 2 cluster is better than 3 clusters (k=3) then optimum number of clusters will be 2 otherwise 3.

Silhouette score Cluster evaluation for KMeans 2 clusters

Silhouette score 2 clusters	0.46577247686580914

Silhouette score Cluster evaluation for KMeans 3 clusters

Silhouette score 3 clusters	0.4007270552751299

KMeans clustering value count of clusters

```
0    71
1    72
2    67
Name: Clus_kmeans, dtype: int64
```

Fig_0.15

Final result add clusters in main data

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

Fig_0.16

Doing Cluster profiling for the 3 clusters as obtained from k-means clustering and taking their mean and generating the data frame for 3 profiles.

Cluster 0: Customer with high spendings, and having less probability of full payment.

Cluster 1: Customer with Medium spendings and having high probability of full payment

Cluster 2: Customer with low spendings and having high probability of full payment

Conclusion | insight: Based on the dataset, 3 clusters can be optimum to do customer segmentation as it is also viable from the k-mean in Full Payer and Revolvers. Because, Non payers are not available in the dataset. Elbow Method is using to find of behave of cluster.

Q1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

Solution-

Recommendation for different promotional Strategies:

1. All result shown in both method Bank should provide more lucrative offers (like cashback /discounts on purchase on different websites/products) to low spending customers to convert them in high spending customers.

2. Almost data was as Bank can also promote some cashback on full payment mechanism which will motivate customers to do more full payments instead of minimum payments.

3. Bank should increase credit limits of full_payments /advance_payments customers to motivate them to do more spending and max spending in single shopping, which may result in EMI and bank can earn on it.

4. Bank can lower the EMI processing fees for high spending customers so that customer can go for EMI options and bank can earn interest on EMI.

5.Using credit Card almost costumer are timely pay one group cluster 1 & cluster 2 are hay ratio so cannulisation are best .

Problem_ 2: CART-RF-ANN

Table of Contents

Contents

Executive Summary

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.

Introduction

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset and analysis of Use CART, RF & ANN and compare the models' performances in train and test sets. The management decides to collect data from the past few years. Target of claimed data all modal are perform as per target.

Data Description

1. Claimed	Claim Status - **Target**
2. Agency Code	Code of tour firm
3. Type	Type of tour insurance firms
4. Channel	Distribution channel of tour insurance agencies
5. Product	Name of the tour insurance products
6. Duration	Duration of the tour
7. Destination	Destination of the tour
8. Sales	Amount of sales of tour insurance policies
9. Commission	The commission received for tour insurance firm
10. Age	Age of insured

Sample of the dataset

	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

Fig_1.0

Conclusion | insight: This is Sample of given data all information is shown in this manner.

Exploratory Data Analysis

```
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
```

Fig_1.2

Conclusion | insight: Nine type of information in this data seat Age, agency_code , Type , Claimed , commission, Channel, Duration ,sales, product Name and Destination data is 234.5 kb and 0 to 2999 row and 10 columns.

Check for missing values in the dataset

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

Fig_1.3

Conclusion | insight: From the above results we can see that there is no missing value present in the dataset.

Check for Duplicate values in the dataset

```
ins_df.duplicated().value_counts()
```

```
False    2861
True      139
dtype: int64
```

```
ins_df.duplicated().sum()
```

```
139
```

```
ins_df.drop_duplicates(inplace=True)
```

```
ins_df.duplicated().value_counts()
```

```
False    2839
dtype: int64
```

Fig_1.4

Conclusion | insight: In this data 139 duplicate value identified but after trite no any duplicate value in this data.

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

Solution-

Univariate analysis

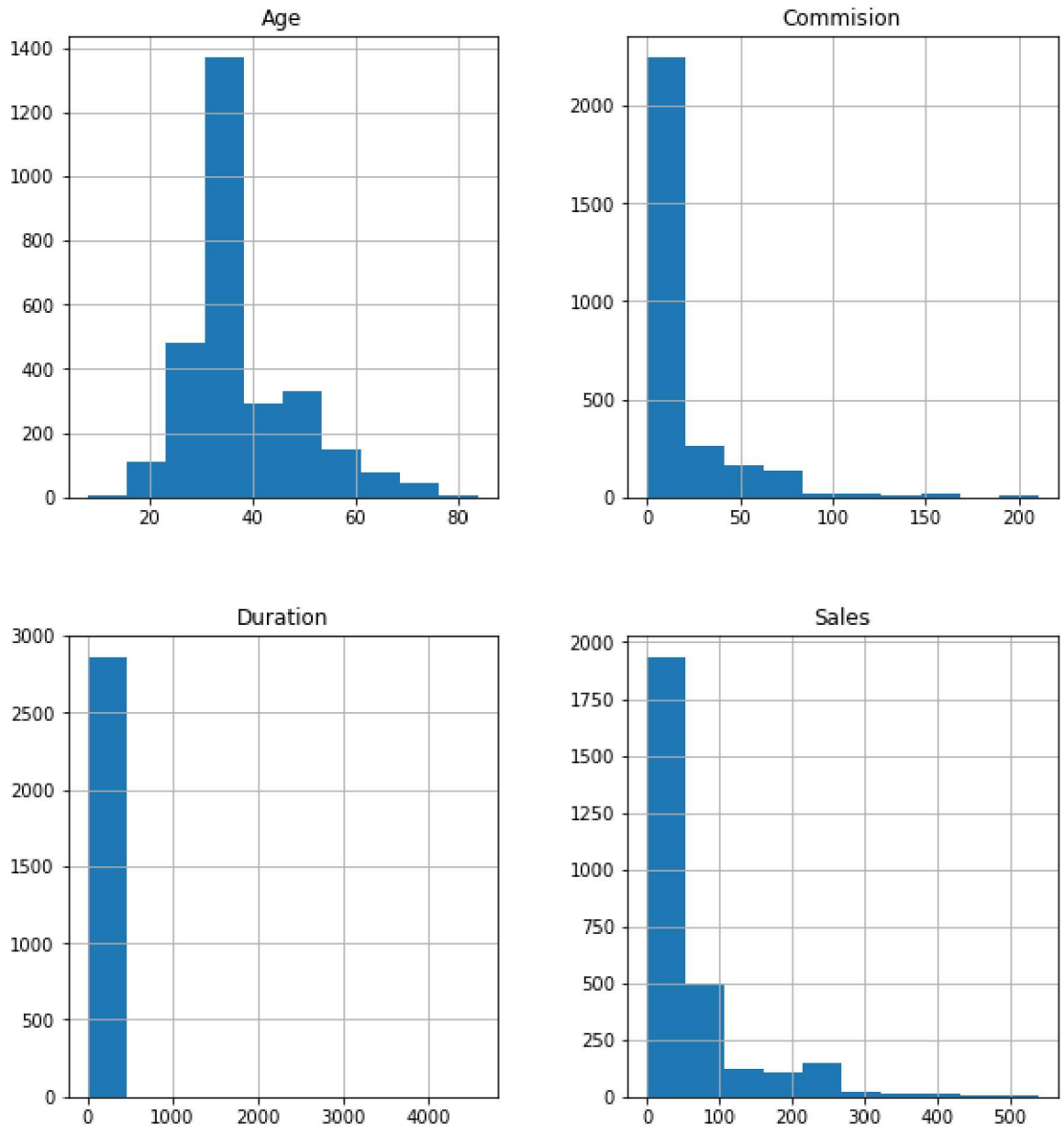

Fig_1.5

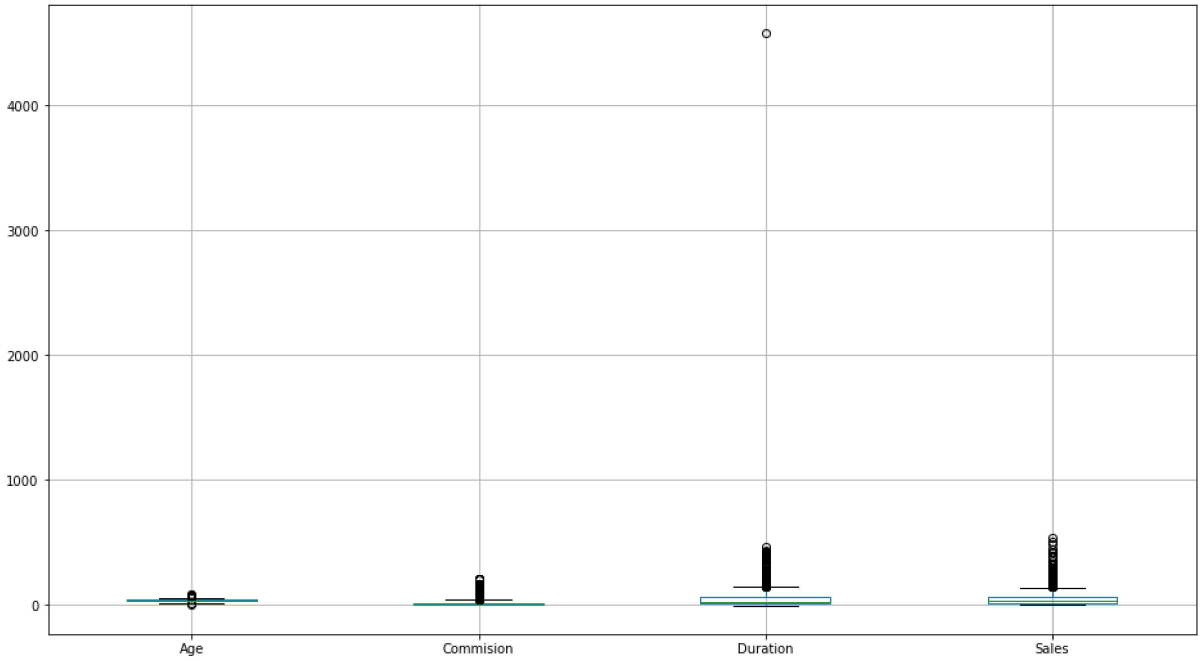

Fig_1.6

Conclusion | insight: All columns data single behaviour show in fig_1.5& fig_1.6

Bivariate Analysis

Column- Claimed graph

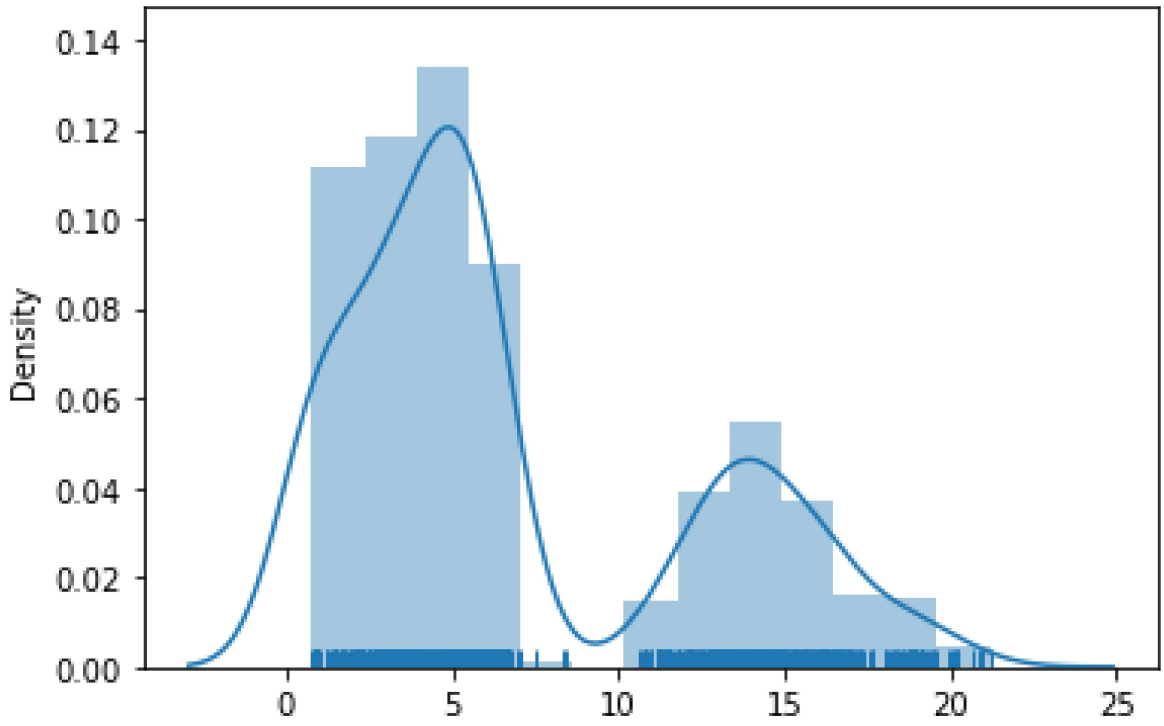

Fig_1.7

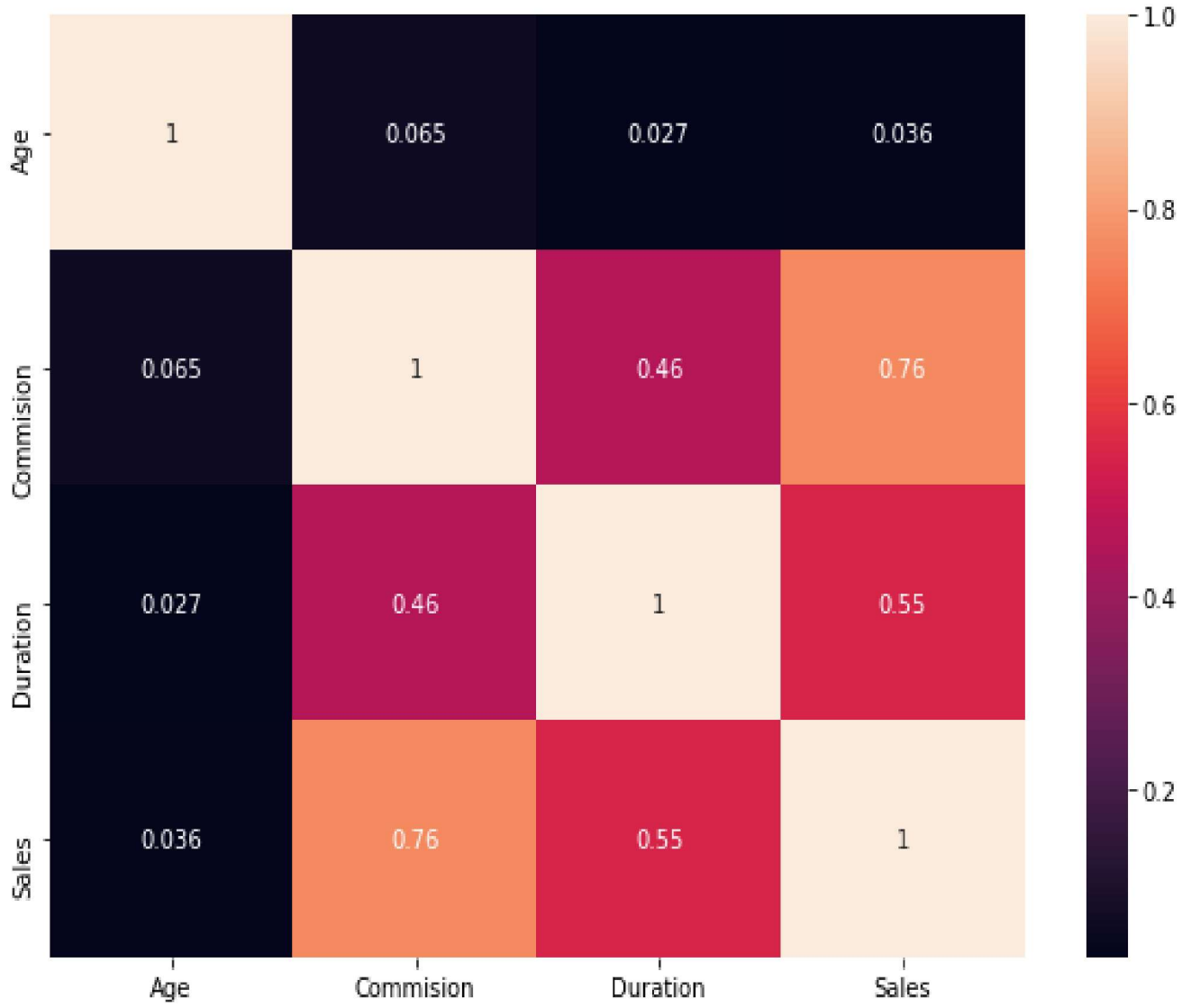

Fig_1.8

Conclusion | insight: Correlation of all two data behaviour in show in fig_1.8

Multivariate Analysis

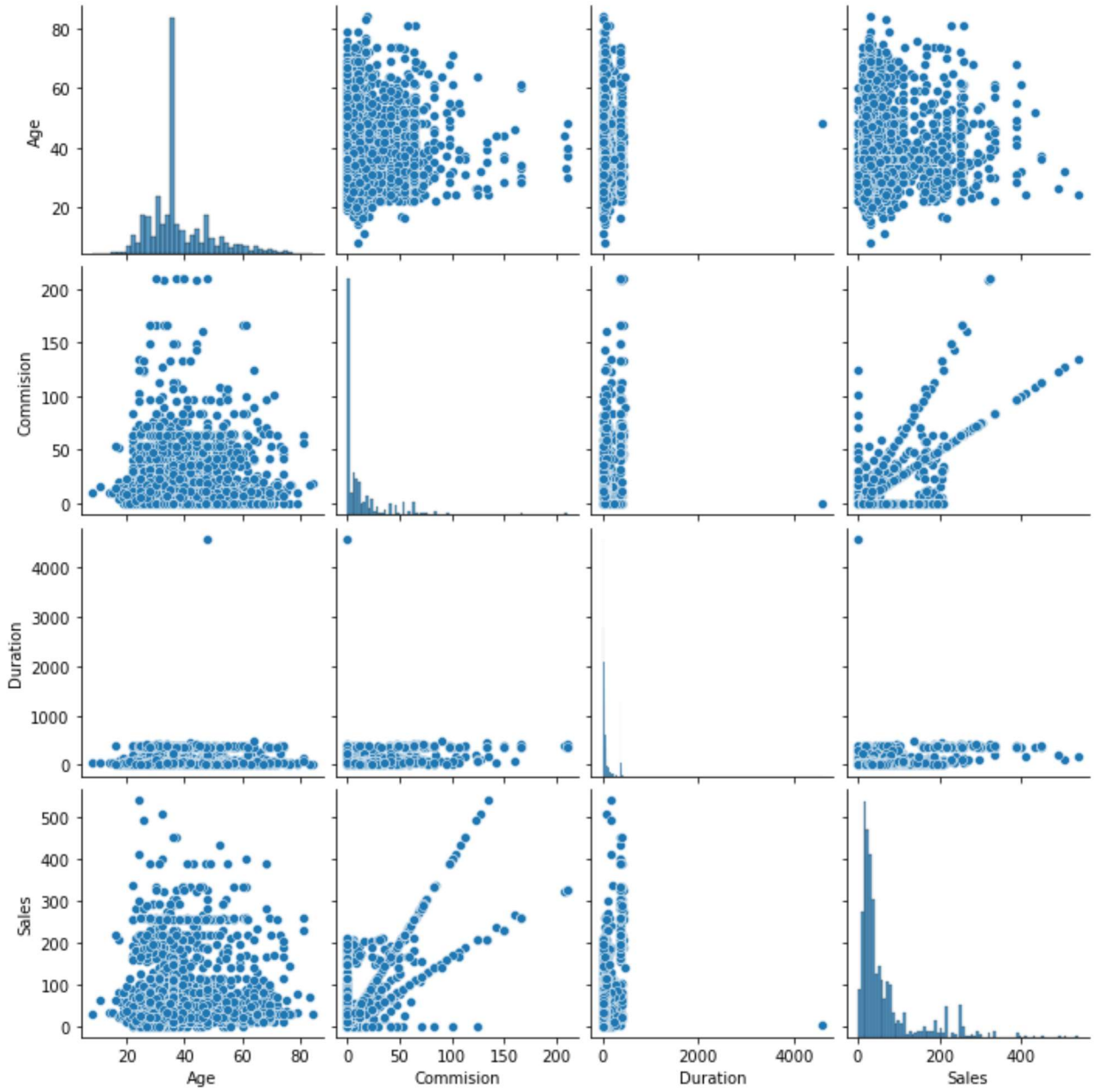

Fig_1.9

Conclusion | insight: All data pair correction and behaviour show in fig_1.9.

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

Solution –

Converted into categorical

Categorised	Claimed	Type	Agency Code	Channel	Product	Destination
0	No'	Airlines	C2B	Online	Customised Plan	'ASIA'
1	Yes	Travel Ag ency	EPX	Offline	Cancellation Plan	'Americas'
2			CWT		Bronze Plan	EUROPE
3			JZI		Silver Plan'	
4					Gold Plan	

Table_1.0

Converted into categorical numerical data

```
Int64Index: 2861 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Age            2861 non-null   int64
 1   Agency_Code    2861 non-null   int8
 2   Type           2861 non-null   int8
 3   Claimed        2861 non-null   int8
 4   Commision      2861 non-null   float64
 5   Channel        2861 non-null   int8
 6   Duration       2861 non-null   int64
 7   Sales          2861 non-null   float64
 8   Product Name   2861 non-null   int8
 9   Destination    2861 non-null   int8
dtypes: float64(2), int64(2), int8(6)
```

Fig_1.10

Categorical sample data

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0	0.70	1	7	2.51	2	0
1	36	2	1	0	0.00	1	34	20.00	2	0
2	39	1	1	0	5.94	1	3	9.90	2	1
3	36	2	1	0	0.00	1	4	26.00	1	0
4	33	3	0	0	6.30	1	53	18.00	0	0

Fig_1.11

Classification model- CART train and test

Grid search for CART

```
GridSearchCV(cv=3, estimator=DecisionTreeClassifier(),
             param_grid={'max_depth': [7, 8, 9, 10],
                         'min_samples_leaf': [15, 20, 25],
                         'min_samples_split': [45, 60, 75]})
```

Fig_1.11b

Performing Grid Search and Cross validation so that best model performance can be obtained.

Best Parameters

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

Fig_1.11c

Using these best parameters, we generate a Decision Tree and build a CART Model and find out the most important variable who will be most decisive for understanding the claim status.

Variable Importance

	Imp
Agency_Code	0.473051
Sales	0.256875
Duration	0.089125
Commision	0.067888
Age	0.067663
Product Name	0.037309
Destination	0.008089
Type	0.000000
Channel	0.000000

Conclusion | insight: After grid search CART Modal data use all data is ready for ROC and AUC, matrix preformed.

Random Forest- train and test

Grid search for RF

```
GridSearchCV(cv=3, estimator=RandomForestClassifier(),
             param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [50, 100],
                         'min_samples_split': [150, 300],
                         'n_estimators': [301, 501]})
```

Performing Grid Search and Cross validation so that best model performance can be obtained.

Best Parameters

```
{'max_depth': 7,
 'max_features': 4,
 'min_samples_leaf': 50,
 'min_samples_split': 150,
 'n_estimators': 301}
```

Min max estimators using in RF Model

Variable Importance

	Imp
Agency_Code	0.346387
Product Name	0.254774
Sales	0.175847
Commision	0.102934
Type	0.047926
Duration	0.043227
Age	0.020765
Destination	0.008140
Channel	0.000000

Conclusion | insight: After grid search RF Modal data use all data is ready for ROC and AUC, matrix preformed.

Artificial Neural Network-train and test

Grid search for ANN

```
GridSearchCV(cv=3, estimator=MLPClassifier(),
             param_grid={'activation': ['logistic', 'relu'],
                         'hidden_layer_sizes': [(100, 100, 100)],
                         'max_iter': [10000], 'solver': ['sgd', 'adam'],
                         'tol': [0.1, 0.01]})
```

Performing Grid Search and Cross validation so that best model performance can be obtained.

Best Parameters

```
{'activation': 'relu',
 'hidden_layer_sizes': (100, 100, 100),
 'max_iter': 10000,
 'solver': 'adam',
 'tol': 0.01}
```

Conclusion | insight: After grid search ANN Modal data use all data is ready for ROC and AUC, matrix preformed.

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

Solution –

1. **CART Modal for Train -**

Confusion matrix –

```
[1199,  160]
[ 262,  381]
```

CART classification –

	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

Cart_train_acc	78%
Cart_train_precision	70%
Cart_train_recall	59%
Cart_train_f1	64%

ROC_AUC score– 0.849

Plot ROC curve

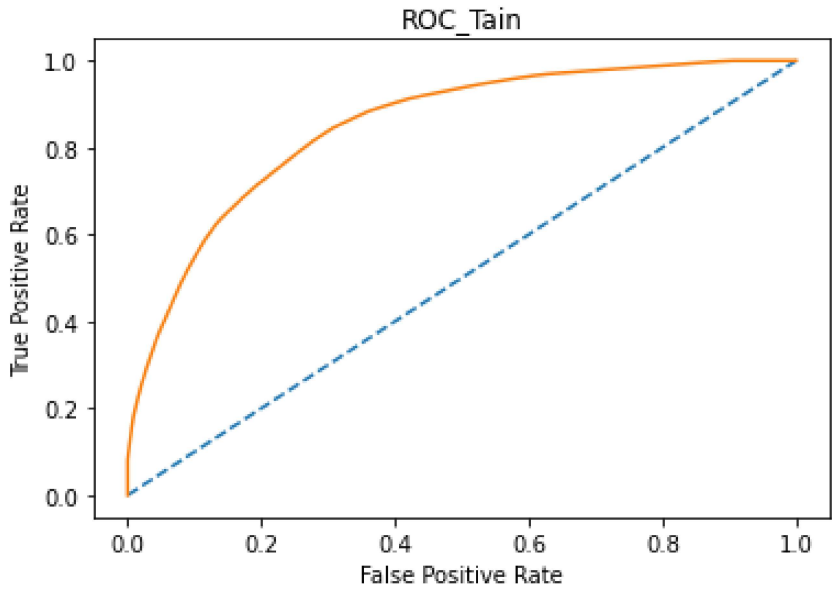

Fig_1.12

CART Modal for Test –

Confusion matrix –

```
[503,  85]
[127, 144]
```

CART classification –

	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	0.71	0.69	0.70	859
weighted avg	0.74	0.75	0.75	859

Cart_train_acc	75%
Cart_train_precision	63%
Cart_train_recall	53%
Cart_train_f1	58%

ROC_AUC score- 0.771

Plot ROC curve

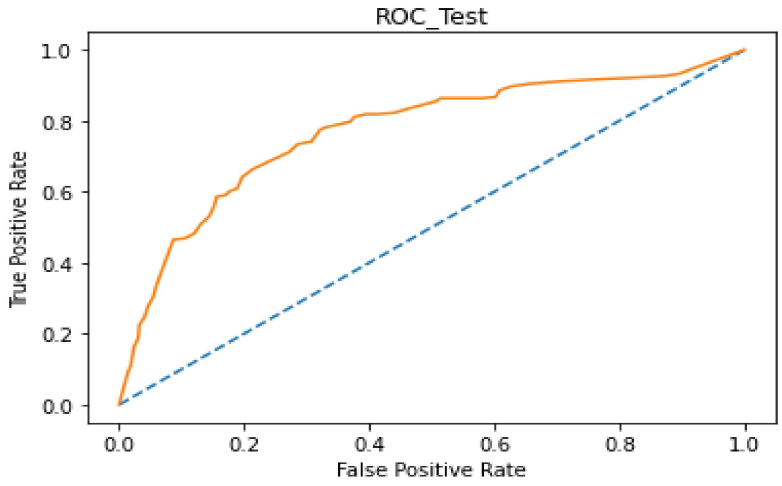

Fig_1.13

Conclusion | insight: In CART Model train and test data and quietly similar result.

2. RF Modal for Train –

Confusion matrix –

```
[1214,  145]
[ 317,  326]
```

RF classification –

	precision	recall	f1-score	support
0	0.79	0.89	0.84	1359
1	0.69	0.51	0.59	643
accuracy			0.77	2002
macro avg	0.74	0.70	0.71	2002
weighted avg	0.76	0.77	0.76	2002

Cart_train_acc	76%
Cart_train_precision	69%
Cart_train_recall	51%
Cart_train_f1	59%

Area under Curve is 0.81

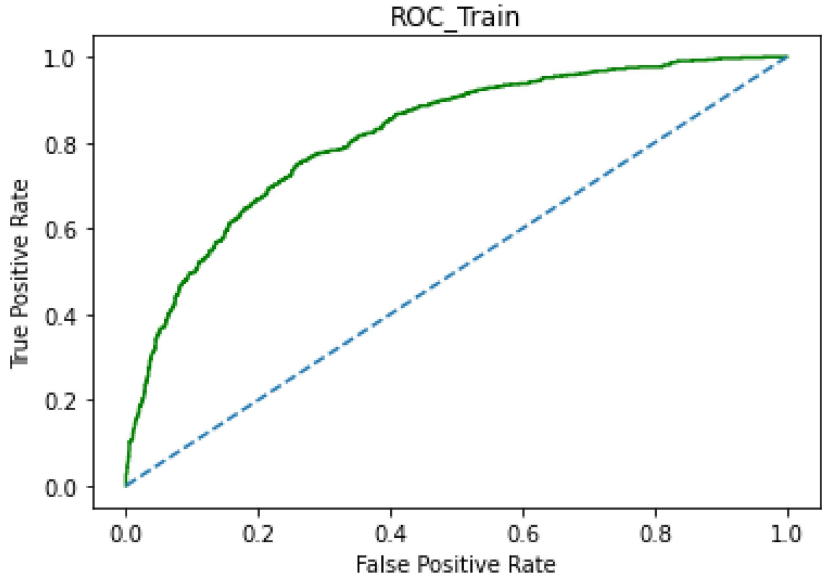

Fig_1.14

RF Modal for Test –

Confusion matrix –

```
[527,  61]
[135, 136]
```

RF classification –

	precision	recall	f1-score	support
0	0.80	0.90	0.84	588
1	0.69	0.50	0.58	271
accuracy			0.77	859
macro avg	0.74	0.70	0.71	859
weighted avg	0.76	0.77	0.76	859

Cart_train_acc	77%
Cart_train_precision	69%
Cart_train_recall	50%
Cart_train_f1	58%

Area under Curve is 0.81

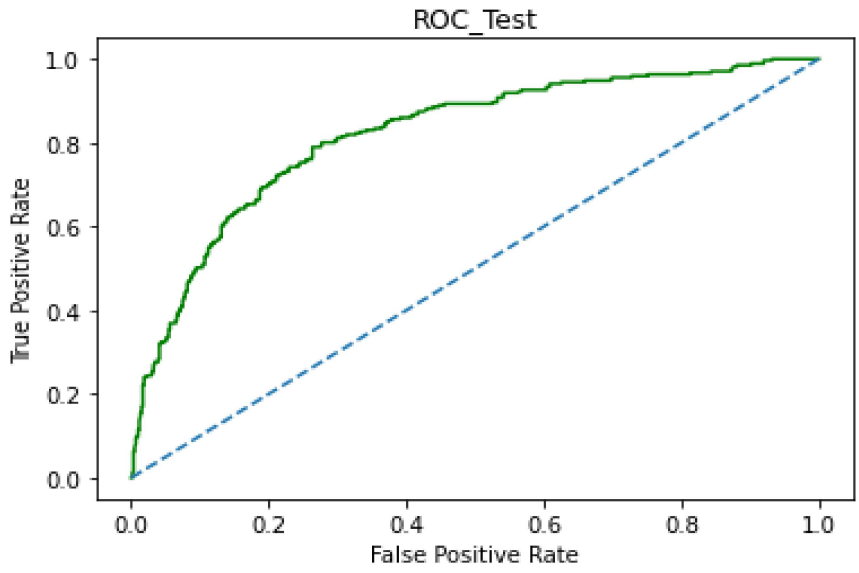

Fig_1.15

Conclusion | insight: In RF Model train and test data and quietly similar result. And best ROC are show .

3. Artificial Neural Network –Train Modal

Confusion matrix –

```
[1172,  187],
[ 272,  371]]
```

ANN classification –

	precision	recall	f1-score	support
0	0.81	0.86	0.84	1359
1	0.66	0.58	0.62	643
accuracy			0.77	2002
macro avg	0.74	0.72	0.73	2002
weighted avg	0.76	0.77	0.77	2002

Cart_train_acc	77%
Cart_train_precision	66%
Cart_train_recall	58%
Cart_train_f1	62%

Area under Curve is 0.7022614057312748

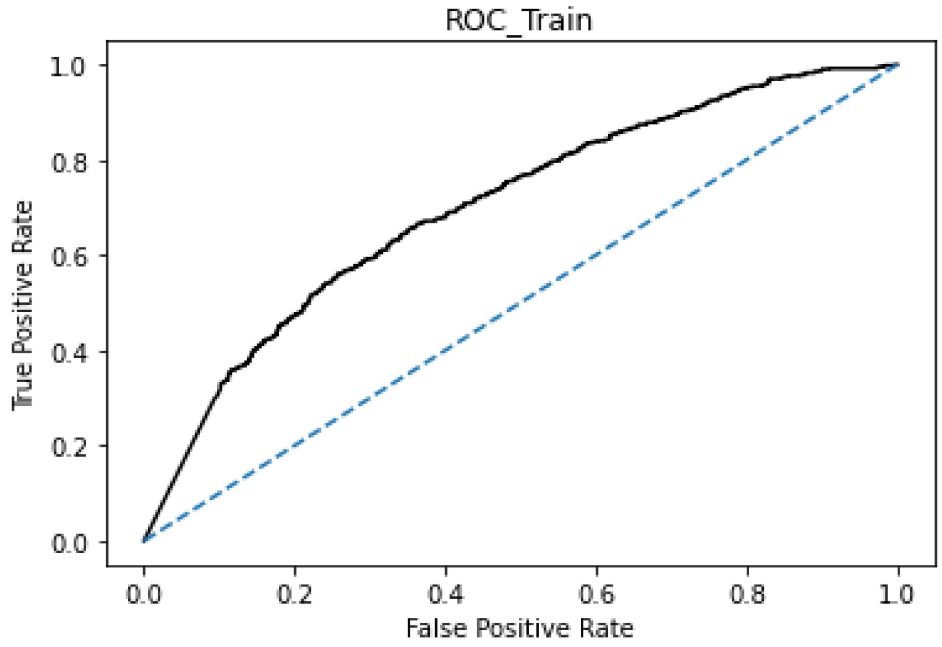

Fig_1.12

Artificial Neural Network –Test Modal

Confusion matrix –

```
[508,  80]
[118, 153]
```

ANN classification –

	precision	recall	f1-score	support
0	0.81	0.86	0.84	588
1	0.66	0.56	0.61	271
accuracy			0.77	859
macro avg	0.73	0.71	0.72	859
weighted avg	0.76	0.77	0.76	859

Cart_train_acc	76%
Cart_train_precision	66%
Cart_train_recall	56%
Cart_train_f1	61%

Area under Curve is 0.6884178025453724

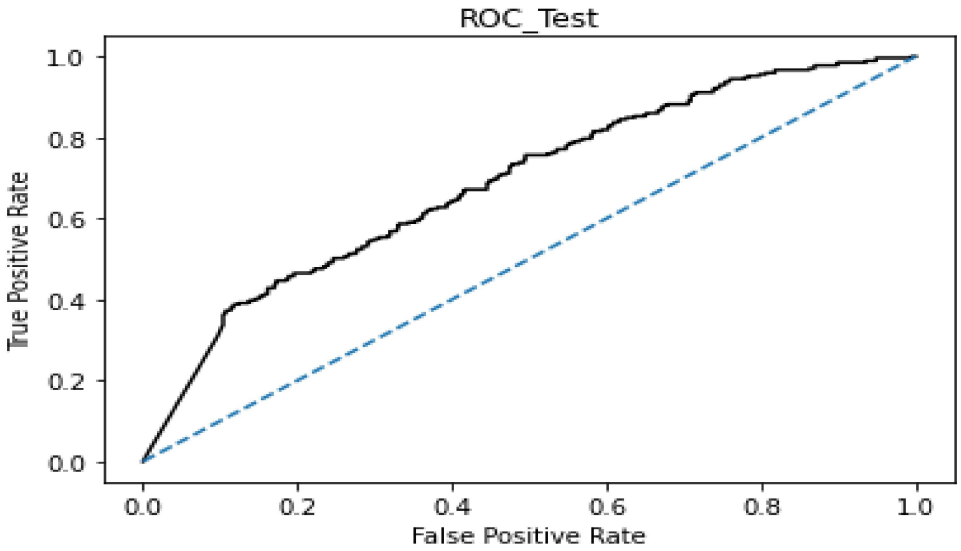

Fig_1.16

Conclusion | insight: In ANN Model train and test data and quietly similar result. And compare to CART and RF model ANN is not perform so good.

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

Solution –

Compare all models on the basis of the performance-

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.79	0.75	0.77	0.77	0.77	0.77
AUC	0.85	0.77	0.82	0.82	0.70	0.69
Recall	0.59	0.53	0.51	0.50	0.58	0.56
Precision	0.70	0.63	0.69	0.69	0.66	0.66
F1 Score	0.64	0.58	0.59	0.58	0.62	0.61

Fig_1.17

ROC Curve for the CART, RF and ANN Train

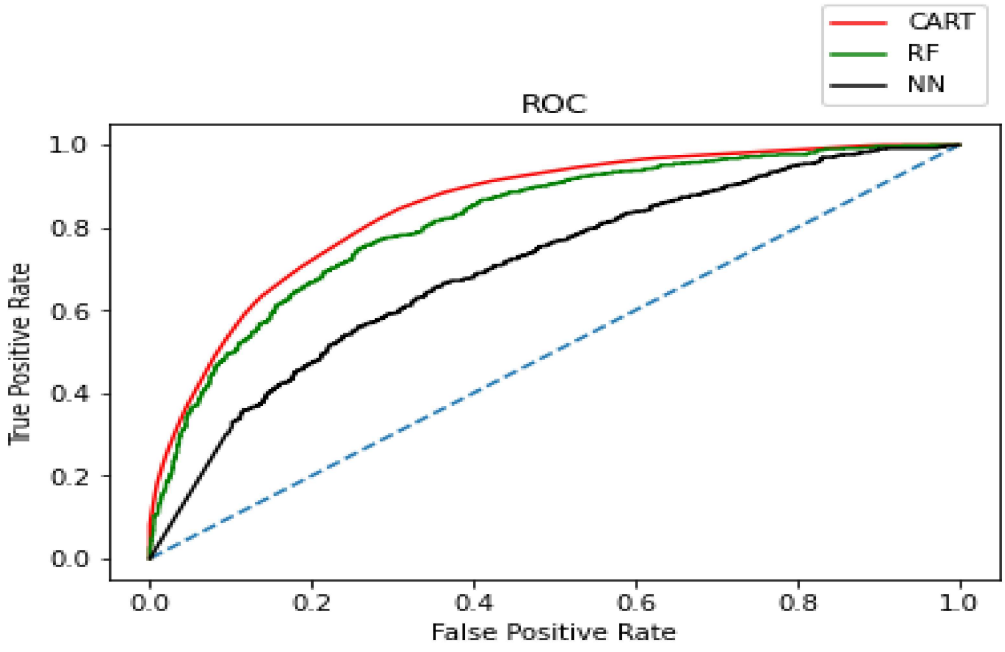

Fig_1.18

ROC Curve for the CART, RF and ANN Test

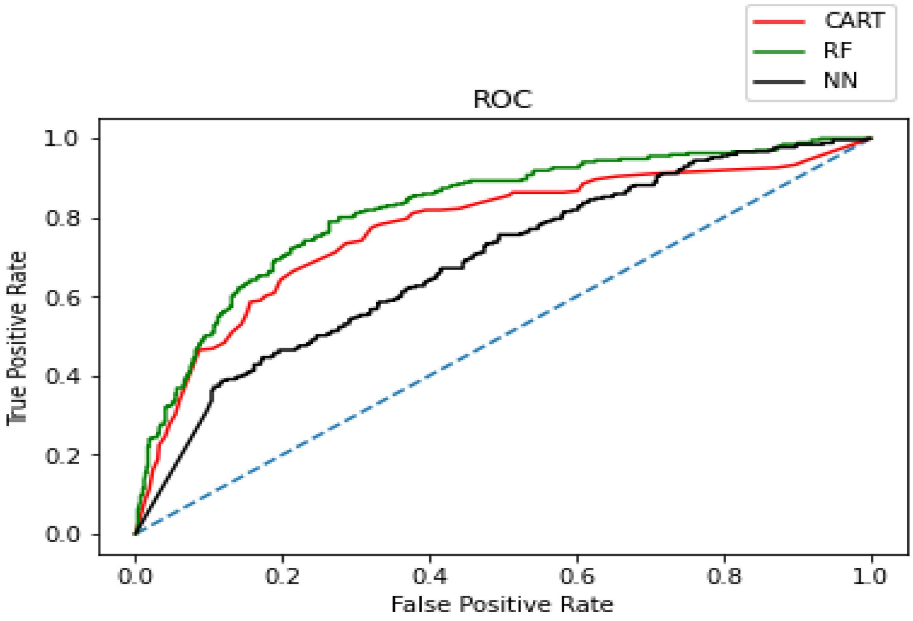

Fig_1.19

Conclusion | insight: As per compare all model RF Model is the best and CART model performs are also good but ANN model performs is low but all model are best performs.

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

Solution –

Conclusion | insight:

1. This is understood by looking at the insurance data by drawing relations between different variables such as day of the incident, time, age group, and associating it with other external information such as location, behaviour patterns, weather information, airline/vehicle etc.

2. Almost need to train the JZI agency resources to pick up sales as they are in bottom, need to run promotional marketing campaign or evaluate if we need to tie up with alternate agency.

3. As per the data 90% of insurance is done by online channel. Other interesting fact is more sales happen via Agency than Airlines and the trend shows the claim are processed more at Airline.

4. Reduce claim handling costs Insights gained from data and AI-powered analytics could expand the boundaries of insurability, extend existing products, and give rise to new risk transfer solutions in areas like a non-damage business interruption and reputational damage.