Data Mining Project

Business report

1.Custering

2. CART-RF-ANN

Student Name: Vivek Bhatia

PGP-DSBA Online Jan_C 2022

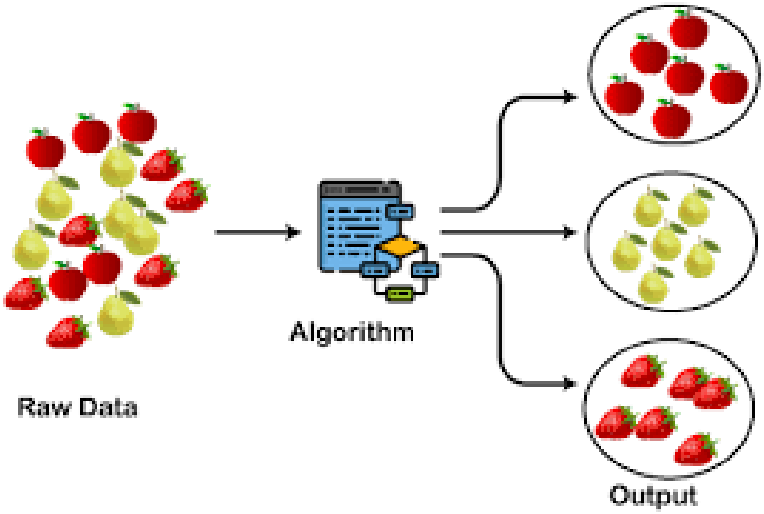
Date: 04/05/2022

Table of Contents

List of figures

List of Pictures

Clustering



Picture 1

1.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.

2.Introduction

The data set provides the various attributes of the bank marketing data for customer like spending, advance payments, probability of full payment, current balance, credit limit, min payment amount and max spending in a single day. This report provides detailed explanation of various problems mentioned in the assignment and its related solution and the inferences made out of that analysis.

3.Data set 1 description

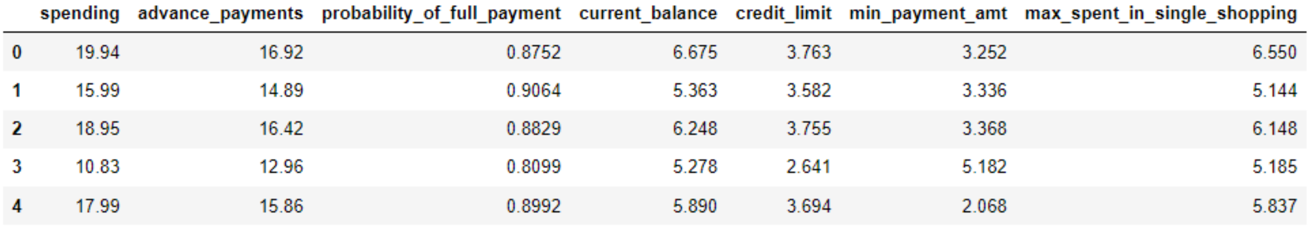
Bank Marketing data was imported into jupyter note book to perform clustering and explain the business implications of performing clustering for this particular case study. The data dictionary is explained below for better understanding of the terminologies used by the bank.

Data Dictionary for Market Segmentation:

- spending: Amount spent by the customer per month (in 1000s)
- advance_payments: Amount paid by the customer in advance by cash (in 100s)
- probability_of_full_payment: Probability of payment done in full by the customer to the bank
- current_balance: Balance amount left in the account to make purchases (in 1000s)
- credit_limit: Limit of the amount in credit card (10000s)
- min_payment_amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)
- max_spent_in_single_shopping: Maximum amount spent in one purchase (in 1000s)

4.Exploratory data analysis

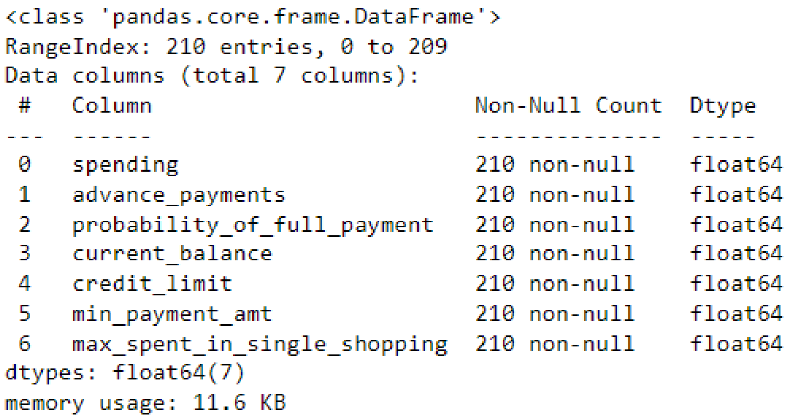
4a. Sample of data set



	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

4b. Check for the type of variables in the data frame

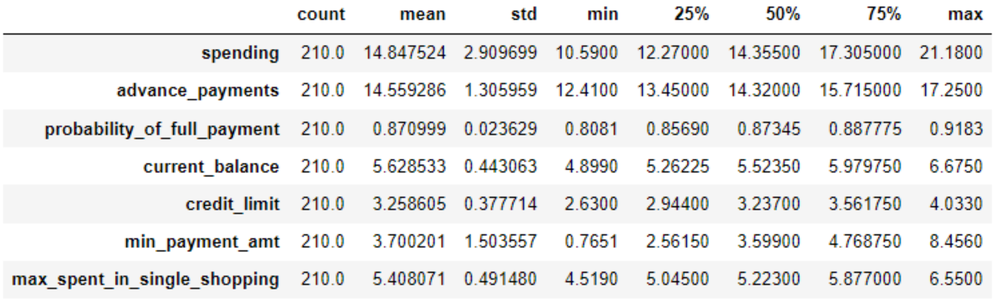
The basic EDA describes the sample data which has 210 rows and 7 columns. There are no null values and all the data types are float. The data describes the spending, advance payments, credit limit and other data related to the credit card.



```
<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
```

5.Question Problem set 1

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



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

By analysing the min, max, mean and standard deviation of the data, we can conclude that there are no abnormal values and we can proceed for further analysis. There are no duplicates in the data set.

1.1.1Univariate analysis:

BY plotting distribution plots and box plots for various attributes of the data we can analyse the data in a detailed manner for the nature of distribution, skew and outliers present.

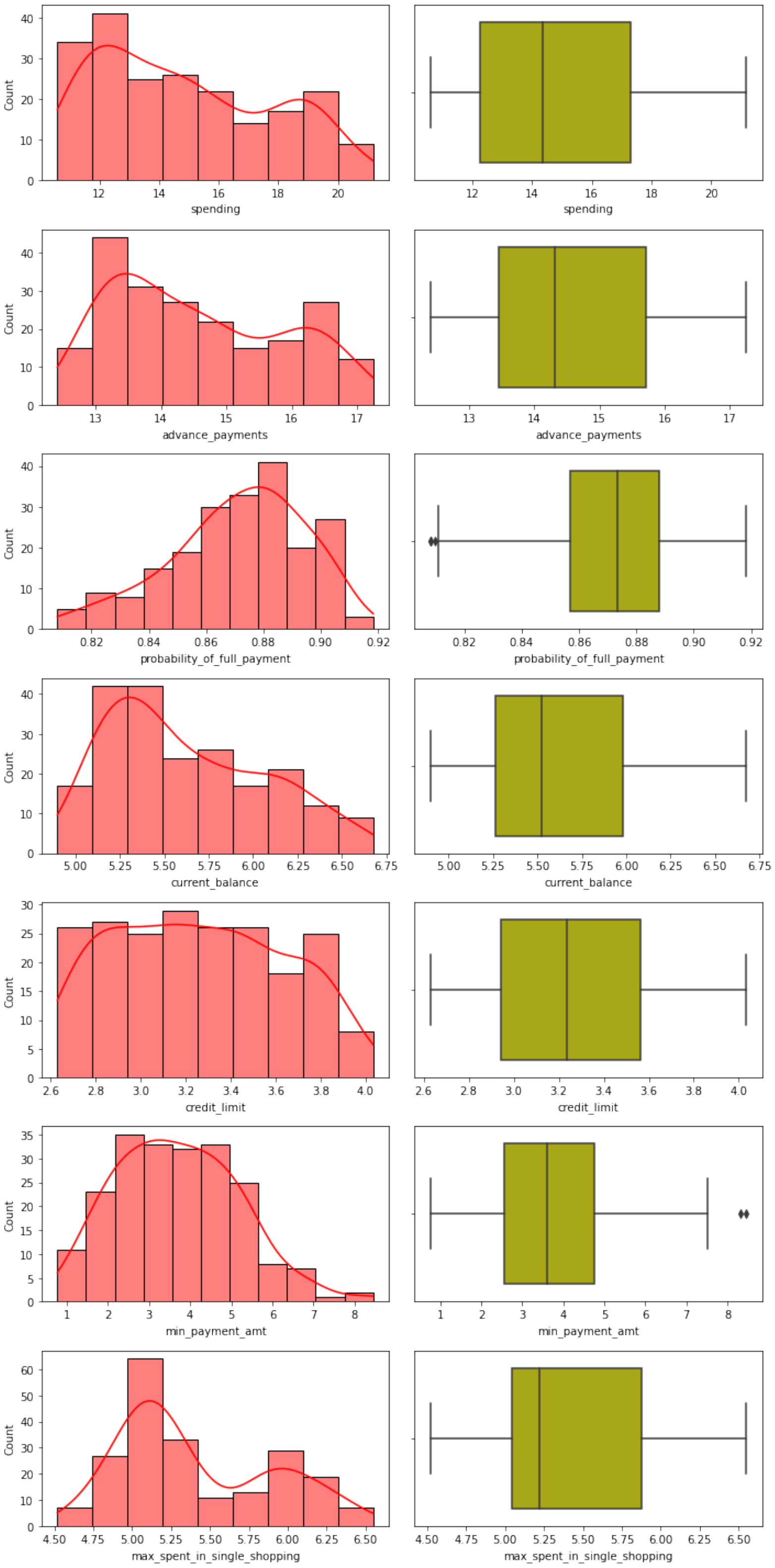
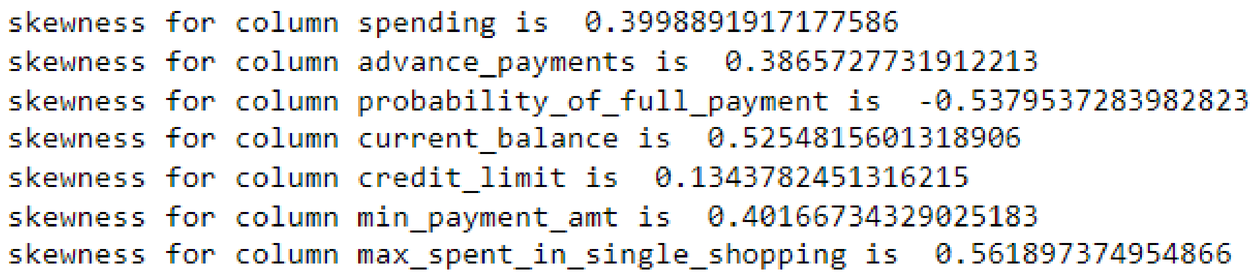


Figure 1

Inferences:



```
skewness for column spending is  0.3998891917177586
skewness for column advance_payments is  0.3865727731912213
skewness for column probability_of_full_payment is  -0.5379537283982823
skewness for column current_balance is  0.5254815601318906
skewness for column credit_limit is  0.1343782451316215
skewness for column min_payment_amt is  0.40166734329025183
skewness for column max_spent_in_single_shopping is  0.561897374954866
```

- **Spending**: Right skewed and no outliers present. Advance Payments: Right skewed and no outliers present.
- **Probability of full payment**: left skewed (Negatively skewed) and has outliers below the min value.
- **Current balance**: Right skewed and no outliers present
- **Credit limit**: Almost normally distributed and no outliers present.
- **Min payment amount**: Almost normally distributed and outliers are present out of the maximum value.
- **Max spent in single shopping**: Right skewed and no outliers present.

1.1.2 Multivariate Analysis:

We use pairplot and heat map to understand the multivariate relationship between various attributes and try to understand the collinearity among data. From the pair plot it is evident that Spending, advance payments, current balance, credit limit and max spent in a single shopping are highly correlated with each other.

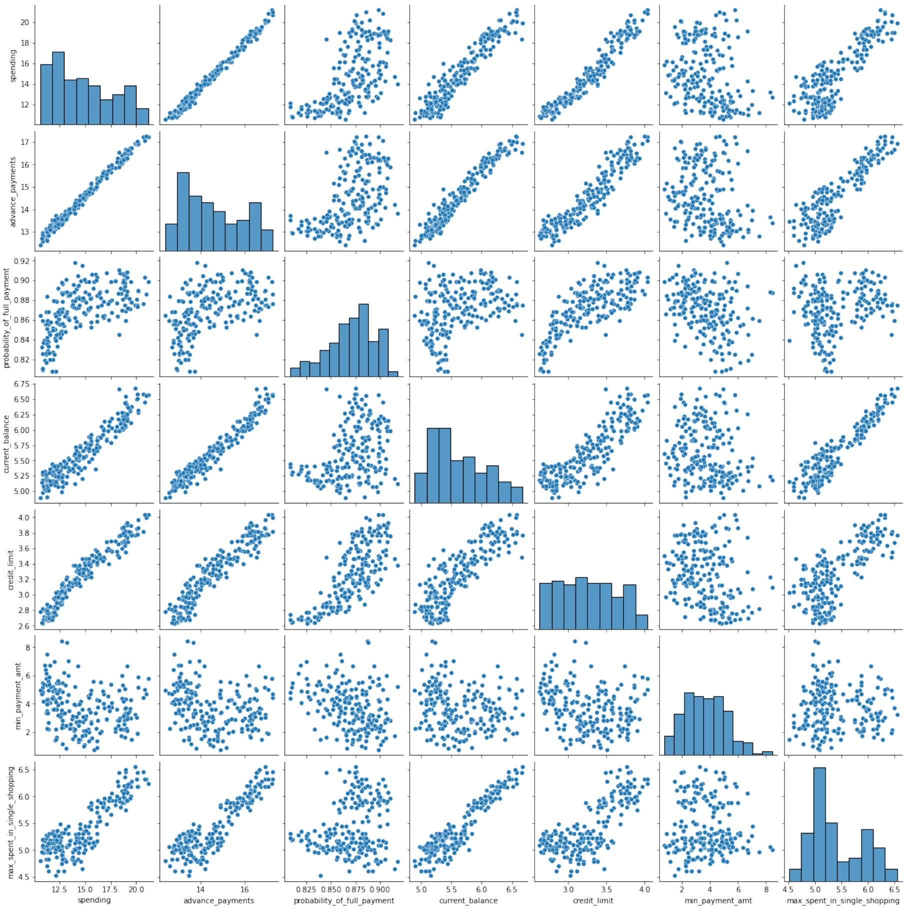


Figure 2

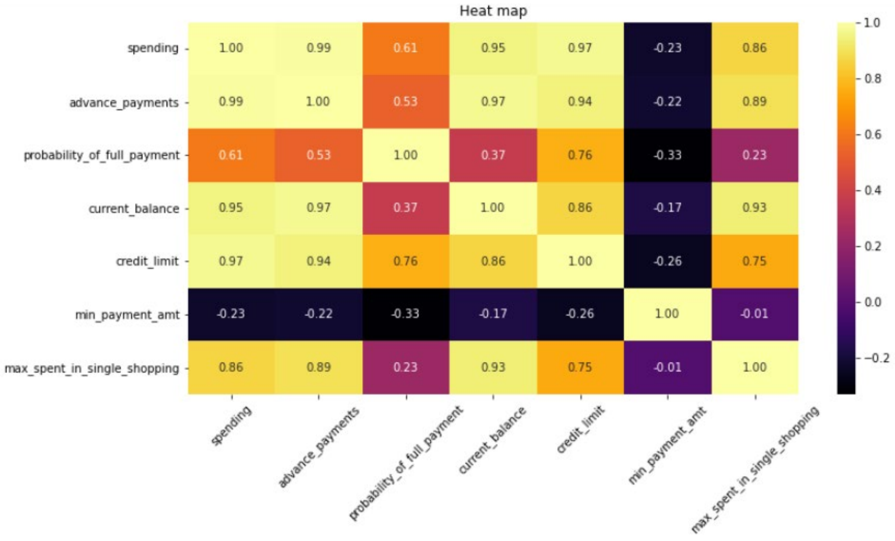


Figure 3

1.1.3 Outlier Treatment

As we have seen outliers are present in the data, we need to treat them before moving ahead for clustering. Post outlier treatment we can see the data as below.

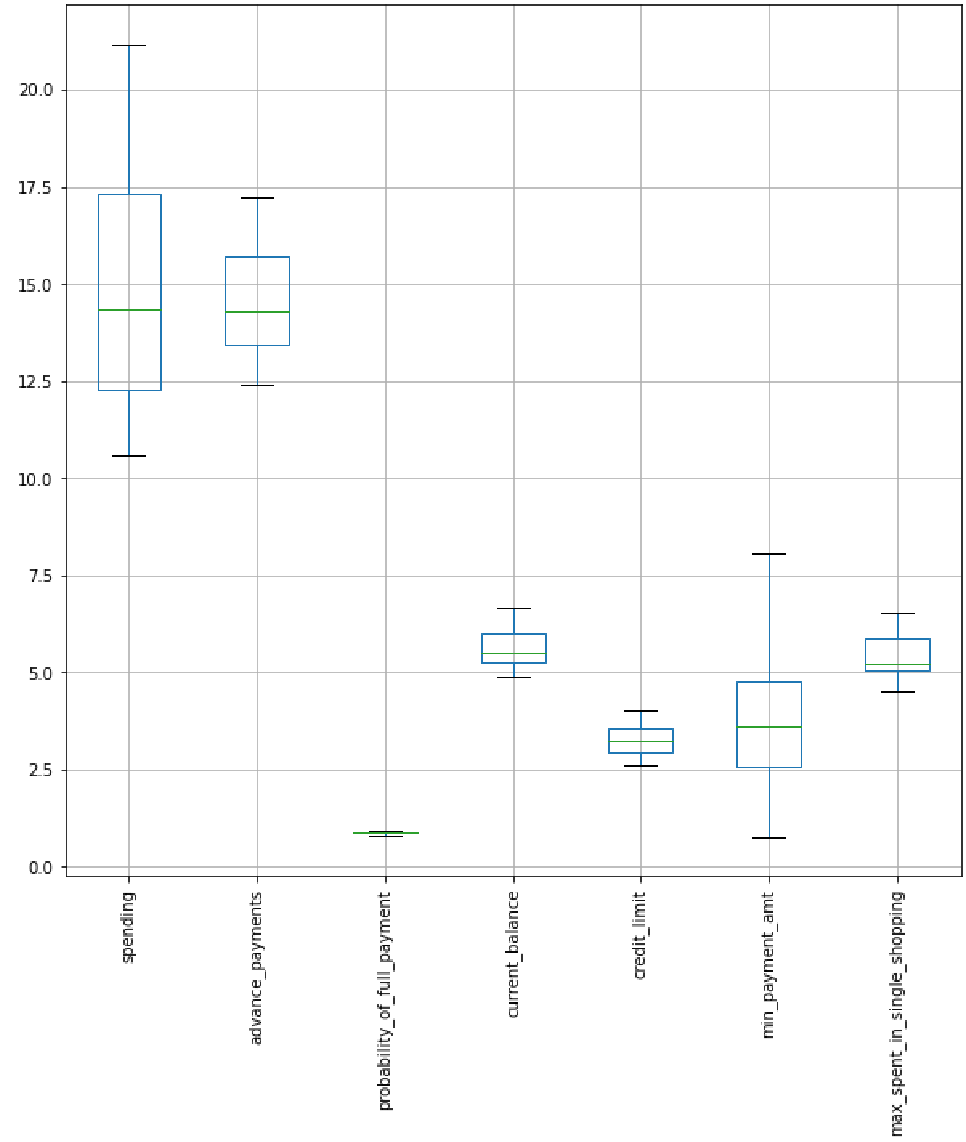
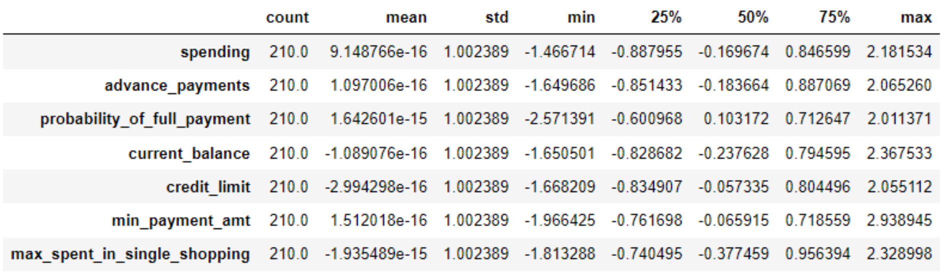


Figure 4

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

As we can see that the mean for spending and advance_payments is considerably different from probability_of_full_payment. Also, the standard deviation of different variables is varying from 2.90 to 0.02 hence scaling will be necessary to bring the variables in the same range. we can use z score, standard scaler or min max scaling method however z scaling is the most commonly used method and we will be using that. Post scaling the description of the data is as below and we can see that the mean and standard deviation are all in the same scale.



	count	mean	std	min	25%	50%	75%	max
spending	210.0	9.148766e-16	1.002389	-1.466714	-0.887955	-0.169674	0.846599	2.181534
advance_payments	210.0	1.097006e-16	1.002389	-1.649686	-0.851433	-0.183664	0.887069	2.065260
probability_of_full_payment	210.0	1.642601e-15	1.002389	-2.571391	-0.600968	0.103172	0.712647	2.011371
current_balance	210.0	-1.089076e-16	1.002389	-1.650501	-0.828682	-0.237628	0.794595	2.367533
credit_limit	210.0	-2.994298e-16	1.002389	-1.668209	-0.834907	-0.057335	0.804496	2.055112
min_payment_amt	210.0	1.512018e-16	1.002389	-1.966425	-0.761698	-0.065915	0.718559	2.938945
max_spent_in_single_shopping	210.0	-1.935489e-15	1.002389	-1.813288	-0.740495	-0.377459	0.956394	2.328998

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

To apply hierarchical clustering, we need to import dendrogram and linkage and use them to create our dendrogram. There are various methods we can use for linkage like average method or ward method and also for criterion we have distance, max clust, monocrit etc. we will be using max-clust criterion for our analysis.

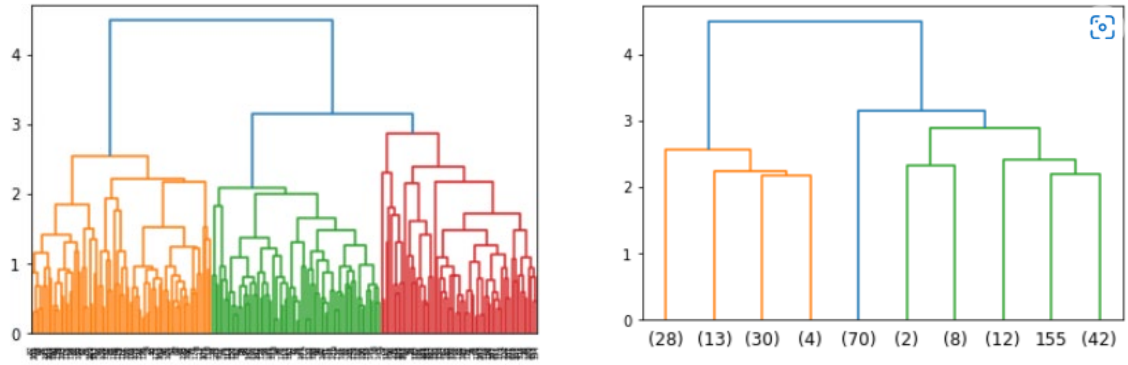
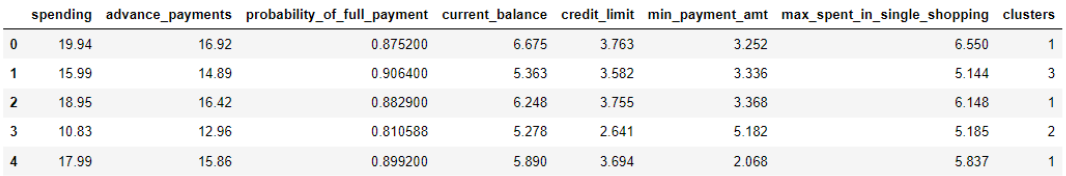


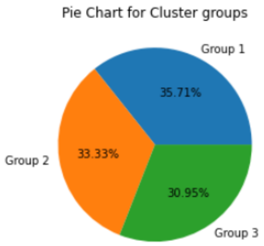
Figure 5

AS seen from the dendrogram before optimising the cluster we have a dendrogram which is broadly grouped into 3 groups identified from different colours like amber, green and red. However, we truncated the dendrogram and have a much better understanding of the clusters.

Once we append the cluster into the original data, we can see the percentage of each group as shown below.



	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	0.875200	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1



As we can see Group 1 is 35.71%, Group 2 is 33.33% and Group 3 is 30.95%

Figure 6

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.

To begin with we import the libraries necessary for clustering and try to understand the wss (within sum of squares) scores and plot an elbow curve. This elbow curve will give us a rough estimate of the optimum number of cluster and we can further refine our inference by checking the silhouette score and silhouette visualiser and come to a pretty good conclusion on what should be our optimum cluster and why?

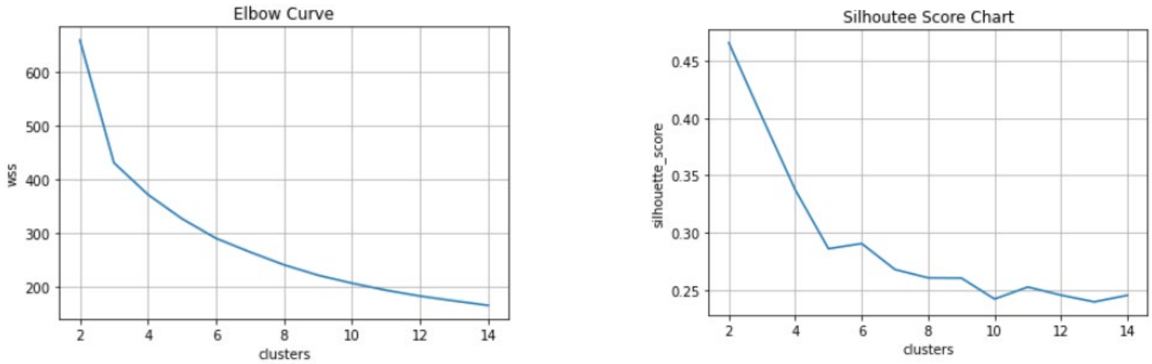


Figure 7

From the elbow curve we can see that our optimal cluster could between 2 to 4 as we see a steep drop in wss from 659.14 at 2 clusters to 371.03 at 4 clusters. Also, based how silhoutte score we can check that the score drops significantly till 4 clusters. To get a more concrete understanding let us visualise the silhoutte visualiser.

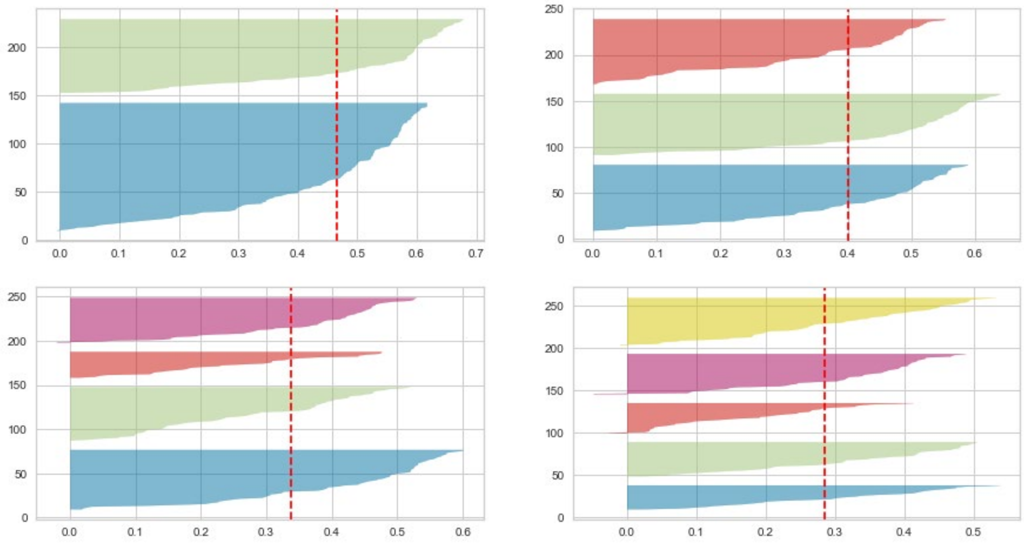


Figure 8

As we can see the silhouette plot for number of clusters at 3 looks optimum based on the below mentioned criteria.

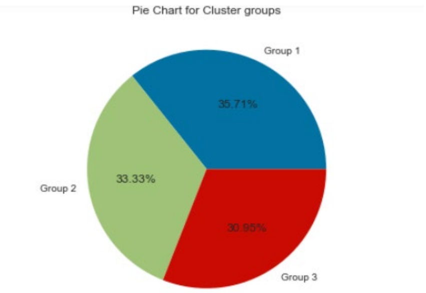
- Average silhouette scores- All the scores are above average silhouette score.
- Size of plot- we can see that there is very less fluctuation in the size of plot when compared with the other plots.
- Uniformity in thickness- we can see that for number of clusters at 3 there is a much better uniformity.

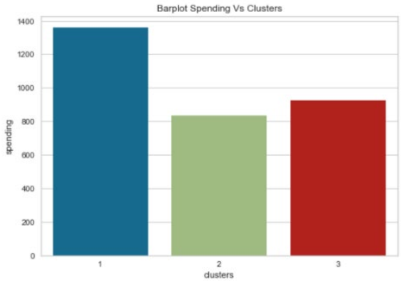
Hence, we can conclude that the best or optimal number of clusters is 3.

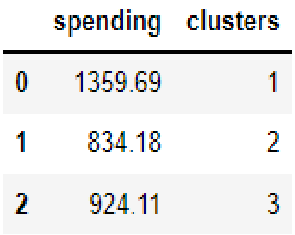
1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

We have already seen that 35.71% of the total customer belong to group 1, 33.33% belong to group 2 and 30.95% belong to group 3. To understand the cluster profiles for the defined clusters we can see the plot of spending with respect to clusters. As spending is highly correlated with the other attributes, we can conclude the same inferences for other variables as we conclude for spending.

Total spending of Group 1 is 1359.69, Group 2 is 834.18 and Group 3 is 924.11.







	spending	clusters
0	1359.69	1
1	834.18	2
2	924.11	3

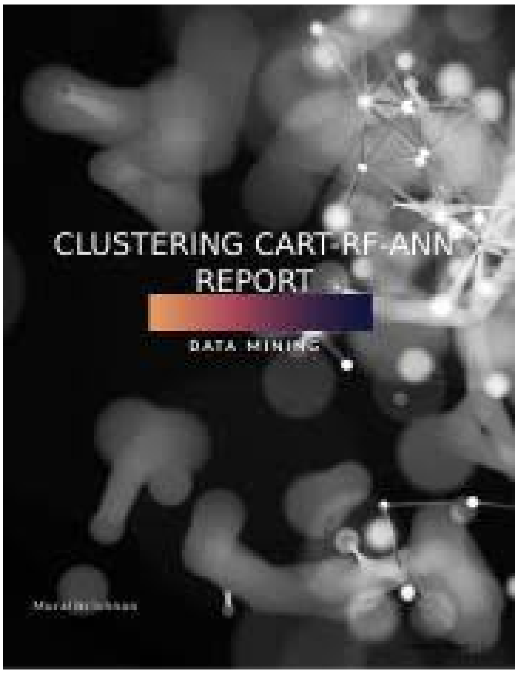
Figure 9

As we can observe Group 1 has the highest spending followed by Group 3 and Group 2. We can now correlate advance payments, current balance, credit limit and max spent in a single shopping with the groups and infer the following points:

- Group 1 belongs to an elite group of customers. They are the most promising customers and should be given special privileges and prime membership options with the best category card like platinum or titanium card.
- Group 3 falls in between and can be encouraged to use more of the bank's services and avail various benefits. They can be counselled via various marketing strategies like email or tele marketing to use their cards and avail more benefits than they are availing at the present moment.
- Group 2 needs real motivation for using their cards and bank can think of giving them special offers and benefits to ensure that these customers increase their card usage. These customers should be made aware of the various benefits they can avail like cash back, points redemptions and promotional offers or zero cost emi options on various shopping merchandise. These initiatives will help them improve the usage and in turn will be profitable for the company.

End of problem set 1-Clustering.

CART-RF-ANN



Picture 2

1.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.

2.Introduction

The data set provides the following Attribute Information:

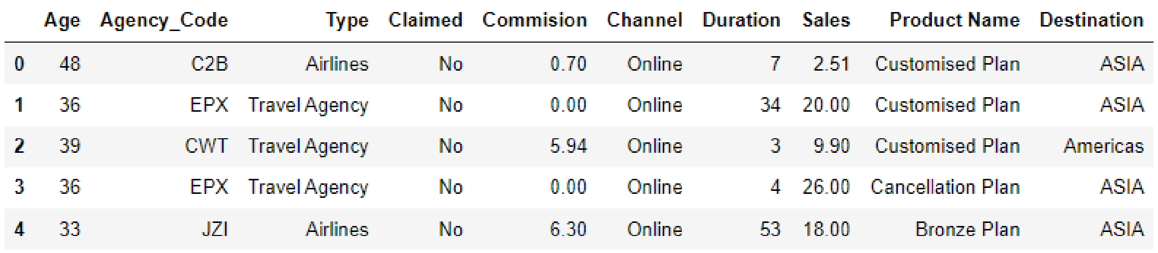
- Target: Claim Status (Claimed)
- Code of tour firm (Agency_Code)
- Type of tour insurance firms (Type)
- Distribution channel of tour insurance agencies (Channel)
- Name of the tour insurance products (Product)
- Duration of the tour (Duration in days)
- Destination of the tour (Destination)
- Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)
- The commission received for tour insurance firm (Commission is in percentage of sales)
- 10.Age of insured (Age)

3.Data set 1 description

Insurance_part_2 data was imported into jupyter note book to analyze and explain the business implications for this particular case study using CART, RF and ANN models and determine which model is best for this particular data set.

4.Exploratory data analysis

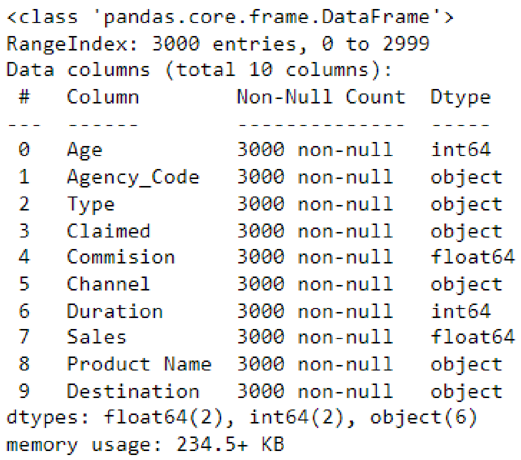
4a. Sample of data set



	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

4b. Check for the type of variables in the data frame

The data set consists of 3000 rows and 10 columns with different variables and the types are shown as below:

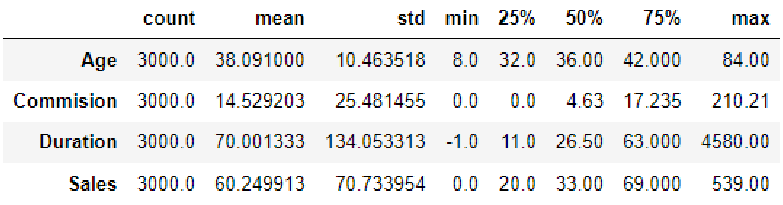


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Age           3000 non-null   int64  
 1   Agency_Code   3000 non-null   object 
 2   Type        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

There are no null values in the data set and the data types are Object, float64 and int64.

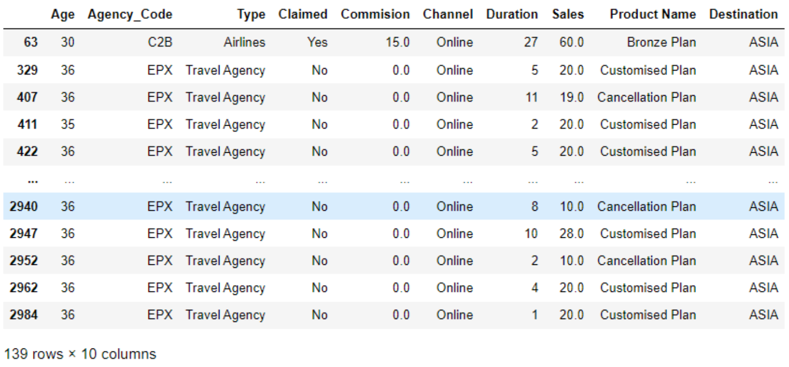
4c. Describe the dataset.

The data description gives the min, max, mean and standard deviation of the attributes. The data looks reasonably good except for Duration which has an abnormal value which is -1 for minimum duration. we will replace it with mean before modelling.



	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

The data also has 139 duplicate rows however we will not drop them as there is no reference to customer name or account id or any other detail which can distinguish the customers data. we can assume here that same set of plans was issued to different customers by same agency, type and for same destination. Hence, we will not drop the duplicate data.



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

139 rows × 10 columns

5.Question Problem set 2

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

Univariate analysis:

To understand and draw conclusions from Univariate analysis divided our data into continuous and categorical for simplicity and plotted them with appropriate histogram, boxplots.

Continuous variables:

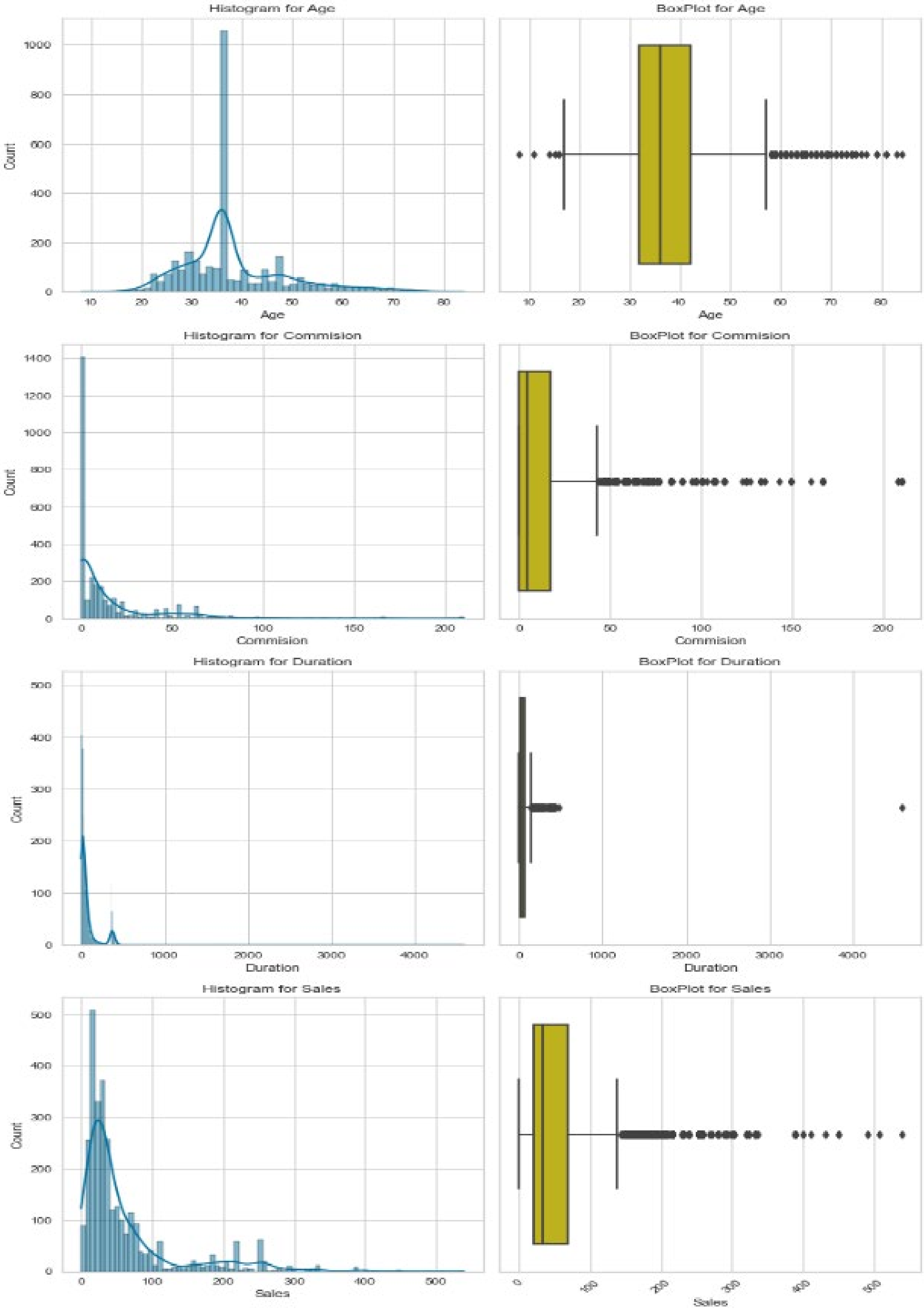


Figure 10

```
skewness for column Age is  1.149712770495169
skewness for column Commision is  3.148857772356885
skewness for column Duration is  13.784681027519602
skewness for column Sales is  2.381148461687274
```

Inferences:

- Age: Normally distributed and has a lot of outliers.
- commission: Right skewed and has lot of outliers.
- Duration: Right skewed and has lot of outliers.
- Sales: Right skewed and has lot of outliers.

Categorical variables:

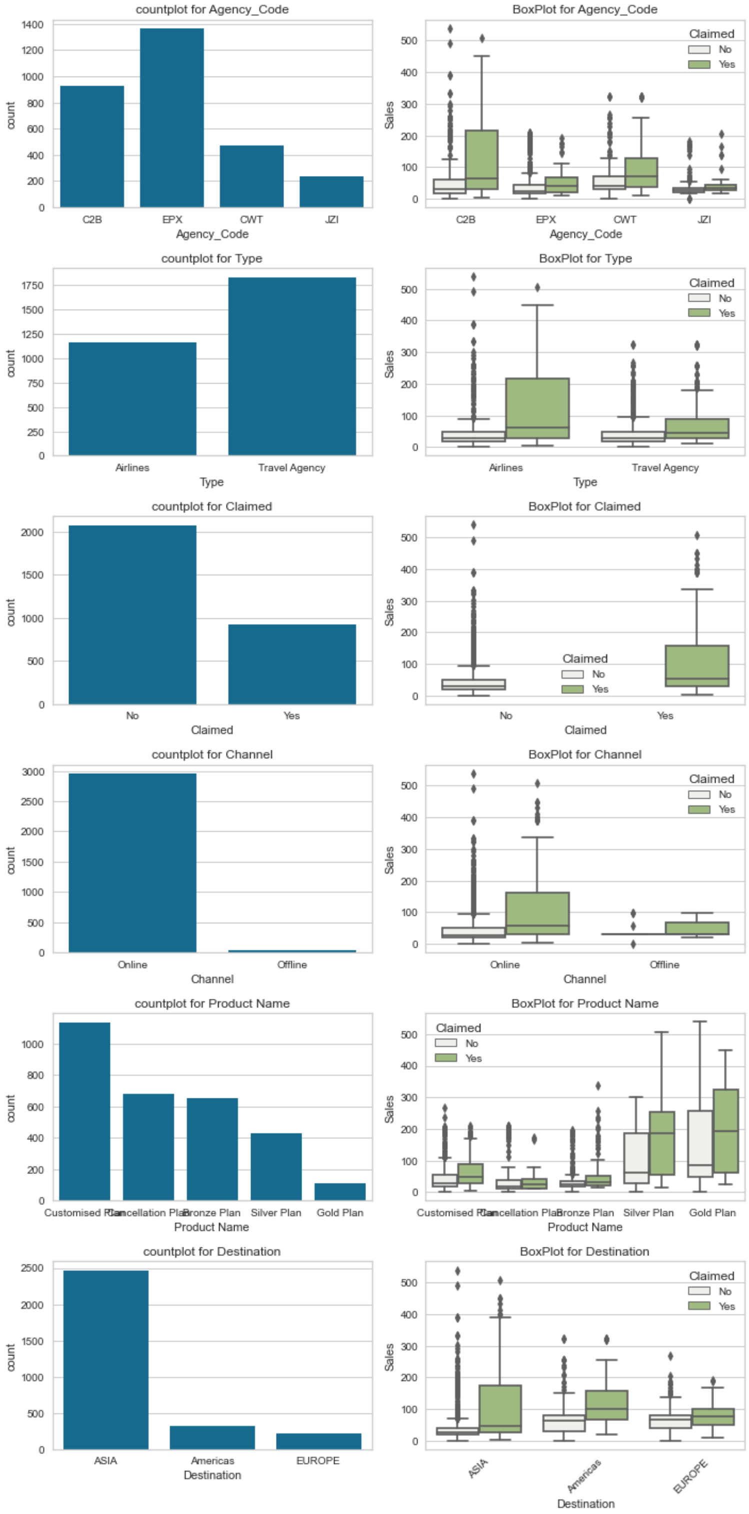


Figure 11

Inferences:

- Agency code: EPX has the maximum count while C2B and CWT has the maximum number of claims settled.
- Type: Travel agencies have more count than airlines however airlines have a greater number of claims settled.
- channel: Online dominates both in count and claim settled.
- Product name: Customized plan is the maximum and Gold plan has the maximum number of claims settled with least customer count.
- Destination: Maximum count is from Asia and Americas has the highest median for claim settled.

Multivariate analysis

To understand and draw conclusions from Multivariate analysis we plot pairplot and heatmap for all the numeric columns and observe the data. Pair plot gives a clear representation of the pairwise correlation between various variables in a data set. Heat map is a value showing chart representing the correlation matrix between different variables

Pair plot

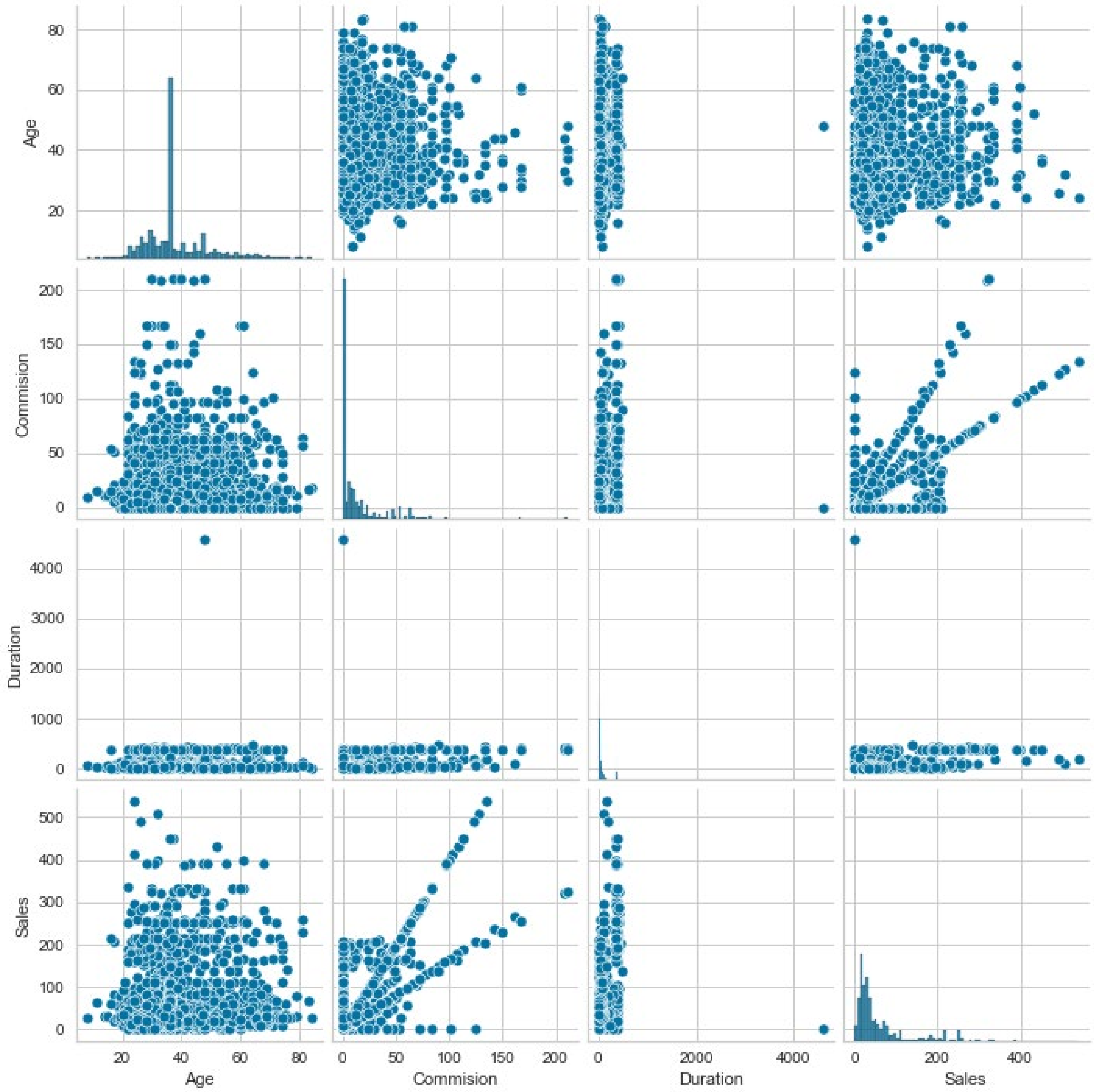


Figure 12

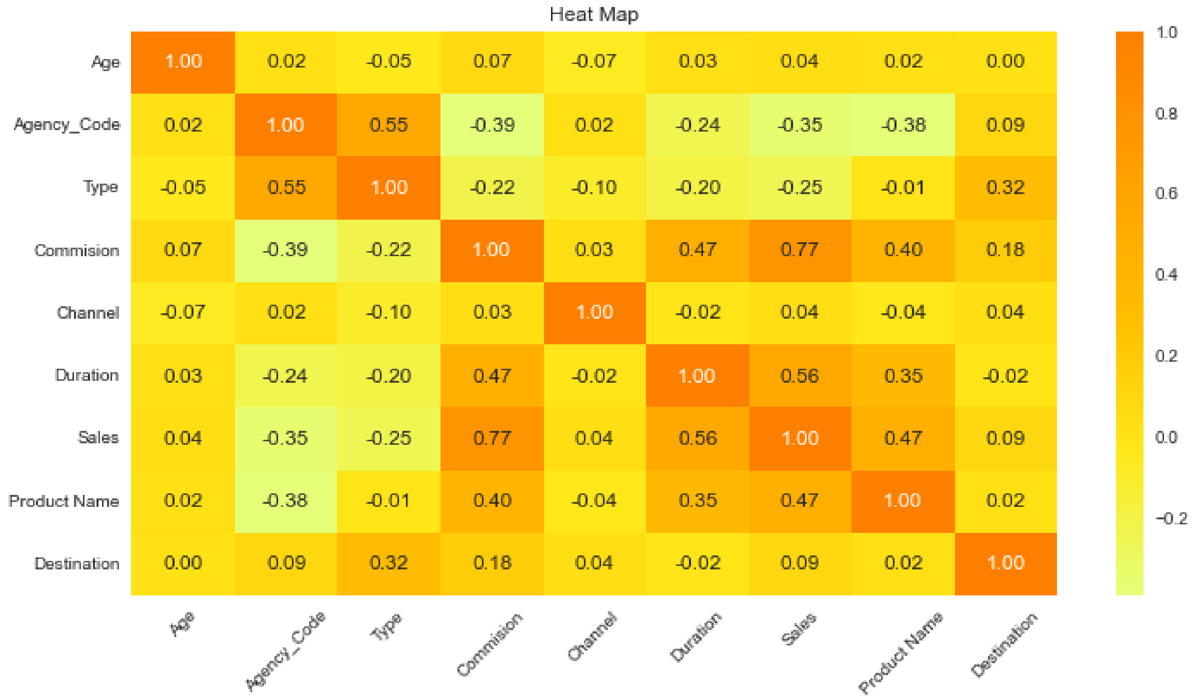
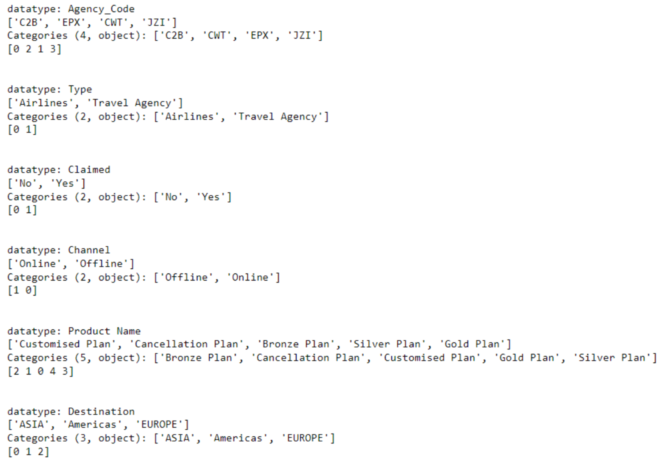


Figure 13

We can see the correlation between sales and commission is 0.77. Other variables do not have much higher correlation with each other.

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

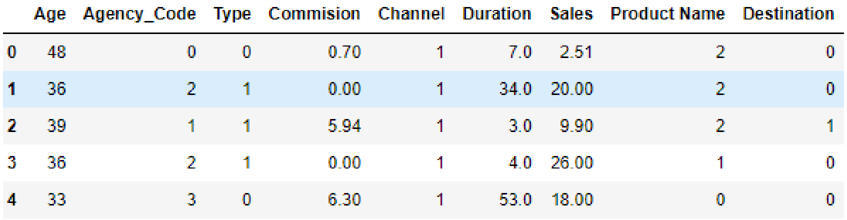
Before we split the data, we need to convert object into categorical data. The split is done on a 70/30 train test ratio as this is considered one of the optimum ratios for splitting the model into train and test.



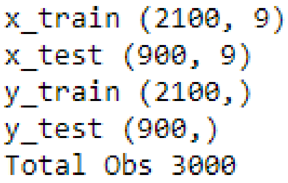
```
datatype: Agency_Code
['C2B', 'EPX', 'CWT', 'JZI']
Categories (4, object): ['C2B', 'CWT', 'EPX', 'JZI']
[0 2 1 3]


datatype: Type
['Airlines', 'Travel Agency']
Categories (2, object): ['Airlines', 'Travel Agency']
[0 1]


datatype: Claimed
['No', 'Yes']
Categories (2, object): ['No', 'Yes']
[0 1]


datatype: Channel
['Online', 'Offline']
Categories (2, object): ['Offline', 'Online']
[1 0]


datatype: Product Name
['Customised Plan', 'Cancellation Plan', 'Bronze Plan', 'Silver Plan', 'Gold Plan']
Categories (5, object): ['Bronze Plan', 'Cancellation Plan', 'Customised Plan', 'Gold Plan', 'Silver Plan']
[2 1 0 4 3]


datatype: Destination
['ASIA', 'Americas', 'EUROPE']
Categories (3, object): ['ASIA', 'Americas', 'EUROPE']
[0 1 2]
```

After converting the data, we will drop the target variable which is “Claimed” in our data set. We split the data into training and test set labelled as “x” and “y” where “x” is data without target variable i.e., claimed and “y” is the target variable.

Sample of the data set after dropping target is shown below. Also, the shape of the train and test model for “x” and “y” is shown below.



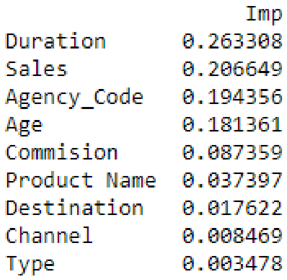
	Age	Agency_Code	Type	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	0	0	0.70	1	7.0	2.51	2	0
1	36	2	1	0.00	1	34.0	20.00	2	0
2	39	1	1	5.94	1	3.0	9.90	2	1
3	36	2	1	0.00	1	4.0	26.00	1	0
4	33	3	0	6.30	1	53.0	18.00	0	0



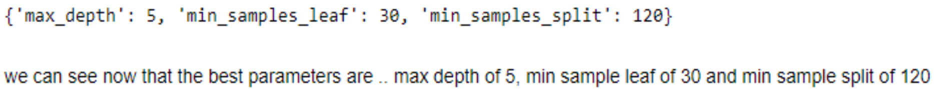
```
x_train (2100, 9)
x_test (900, 9)
y_train (2100,)
y_test (900,)
Total Obs 3000
```

Once the split is done, we will build Decision tree (CART) model, Random Forest classifier (RFCL) model and Artificial neural network (ANN) model.

CART model: In CART model we use Decision tree classifier and fit our train and test data into the dt_model (Decision Tree model). We can get a list of important features as per the weightage they carry and will influence the decision tree. Without pruning the data, the decision tree overgrows and hence we need to do a grid search to get optimum parameters for our decision tree model. By trying various combinations of the grid parameters, we conclude the following best parameters for our decision tree model.



```
                   Imp
Duration      0.263308
Sales         0.206649
Agency_Code   0.194356
Age           0.181361
Commision     0.087359
Product Name  0.037397
Destination   0.017622
Channel       0.008469
Type          0.003478
```

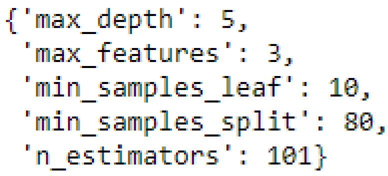


```
{'max_depth': 5, 'min_samples_leaf': 30, 'min_samples_split': 120}
```

we can see now that the best parameters are .. max depth of 5, min sample leaf of 30 and min sample split of 120

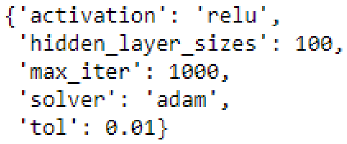
- Duration is the most important feature of our model followed by sales, Agency code, Age and the least important feature is Type.
- The best parameters are max depth: 5, min sample leaf: 30, min sample split: 120.
  Once we have established the model, we are ready to check the accuracy of the train and test samples.

RFCL model: RFCL algorithm output are set of decision trees that work according to the output. Hence, it can be considered as an extended version of decision tree model. It uses ensemble method and the best parameters for the RFCL model is as below:



```
{'max_depth': 5,
 'max_features': 3,
 'min_samples_leaf': 10,
 'min_samples_split': 80,
 'n_estimators': 101}
```

ANN model: ANN model uses hidden layers and an activation function algorithm to determine the output based on the input data. Scaling is important for ANN and we used standard scaler to scale our train and test data. The best parameters for ANN model were observed to be the following:



```
{'activation': 'relu',
 'hidden_layer_sizes': 100,
 'max_iter': 1000,
 'solver': 'adam',
 'tol': 0.01}
```

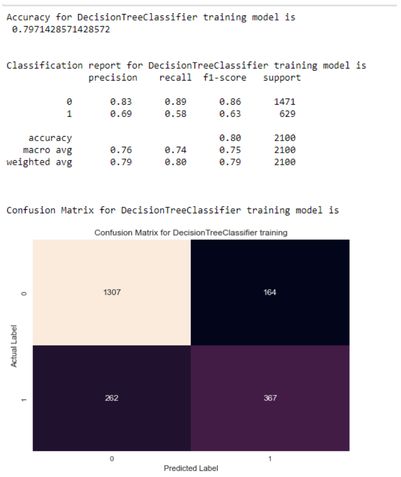
With the inputs for our MLP classifier we had 15 iterations before the epochs stopped.

Inputs: hidden_layer_sizes=100, max_iter=1000, solver='adam’, verbose=True, random_state=21, tol=0.01

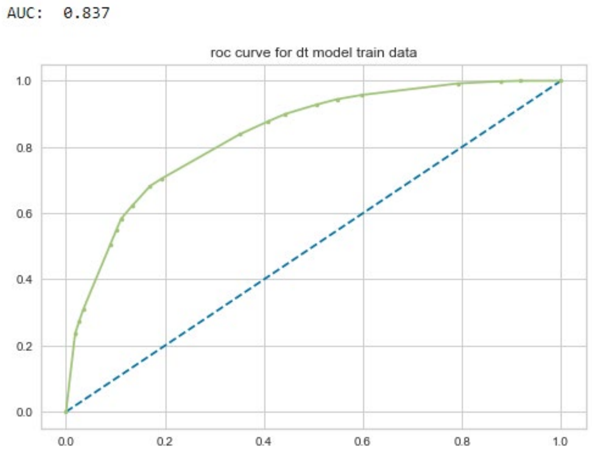
```
Iteration 1, loss = 0.61364553
Iteration 2, loss = 0.55129553
Iteration 3, loss = 0.51554654
Iteration 4, loss = 0.49724773
Iteration 5, loss = 0.48744163
Iteration 6, loss = 0.48234826
Iteration 7, loss = 0.47921250
Iteration 8, loss = 0.47782275
Iteration 9, loss = 0.47660548
Iteration 10, loss = 0.47583911
Iteration 11, loss = 0.47496588
Iteration 12, loss = 0.47429522
Iteration 13, loss = 0.47375464
Iteration 14, loss = 0.47321975
Iteration 15, loss = 0.47258170
Training loss did not improve more than tol=0.010000 for 10 consecutive epochs. Stopping.
```

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

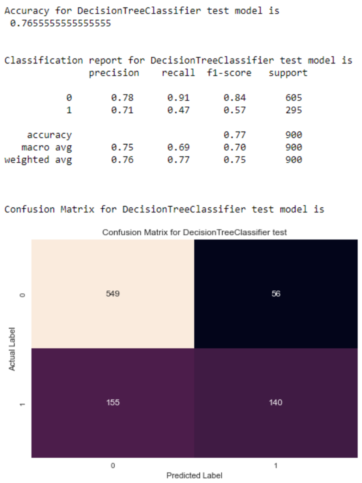
CART model:



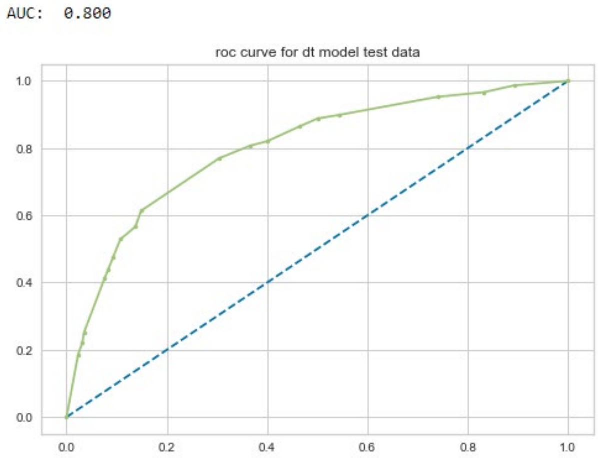
```
Accuracy for DecisionTreeClassifier training model is
 0.7971428571428572


Classification report for DecisionTreeClassifier training model is
              precision    recall  f1-score   support

           0       0.83      0.89      0.86      1471
           1       0.69      0.58      0.63       629

    accuracy                           0.80      2100
   macro avg       0.76      0.74      0.75      2100
weighted avg       0.79      0.80      0.79      2100


Confusion Matrix for DecisionTreeClassifier training model is
```



```
AUC:  0.837
```



```
Accuracy for DecisionTreeClassifier test model is
 0.7655555555555555


Classification report for DecisionTreeClassifier test model is
              precision    recall  f1-score   support

           0       0.78      0.91      0.84       605
           1       0.71      0.47      0.57       295

    accuracy                           0.77       900
   macro avg       0.75      0.69      0.70       900
weighted avg       0.76      0.77      0.75       900


Confusion Matrix for DecisionTreeClassifier test model is
```



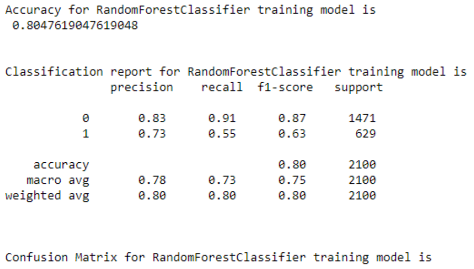
```
AUC:  0.800
```

Figure 14

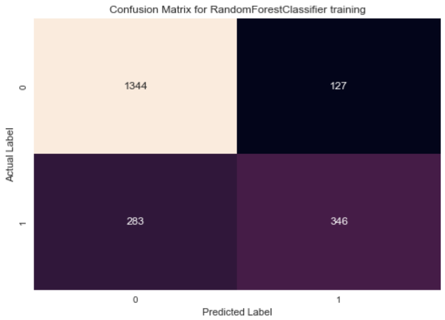
Inferences:

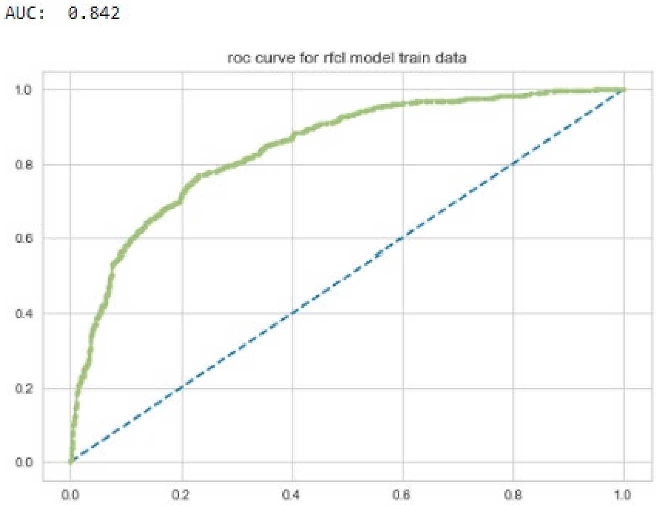
- Train data-Accuracy of train model is 80.0, recall is 0.58, precision is 0.69 and auc is 83.7.
- Test data- Accuracy is 77.0, recall is 0.47, precision is 0.71 and auc is 80.0.
- Although the test data is performing close to the train data, we definitely have a room here for improvement in our training data to improve the accuracy further.

RFCL model:

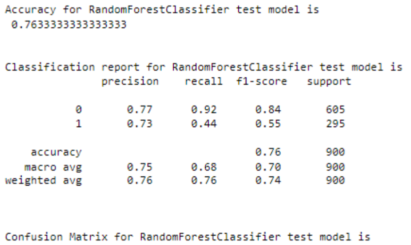


```
Accuracy for RandomForestClassifier training model is
 0.8047619047619048

Classification report for RandomForestClassifier training model is
               precision    recall  f1-score   support

           0       0.83      0.91      0.87      1471
           1       0.73      0.55      0.63       629

    accuracy                           0.80      2100
   macro avg       0.78      0.73      0.75      2100
weighted avg       0.80      0.80      0.80      2100


Confusion Matrix for RandomForestClassifier training model is
```

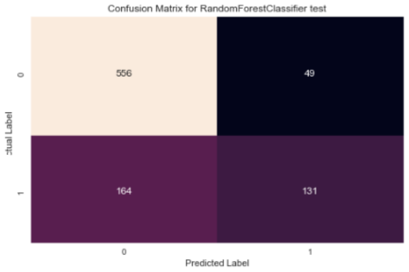


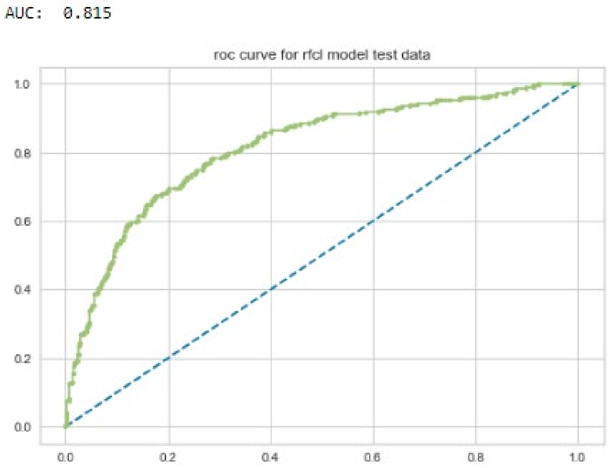


AUC: 0.842



```
Accuracy for RandomForestClassifier test model is
 0.7633333333333333

Classification report for RandomForestClassifier test model is
              precision    recall  f1-score   support

           0       0.77      0.92      0.84       605
           1       0.73      0.44      0.55       295

    accuracy                           0.76       900
   macro avg       0.75      0.68      0.70       900
weighted avg       0.76      0.76      0.74       900


Confusion Matrix for RandomForestClassifier test model is
```





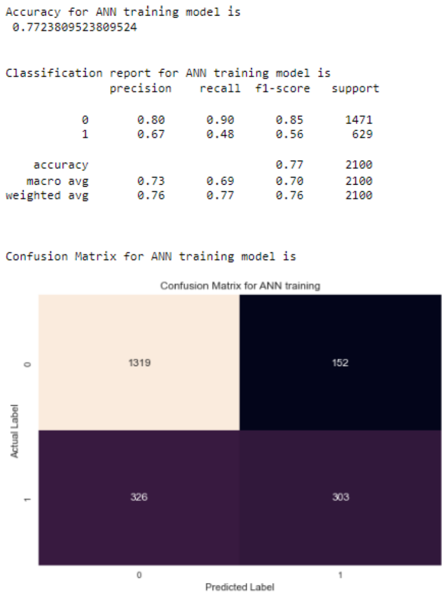
AUC: 0.815

Figure 15

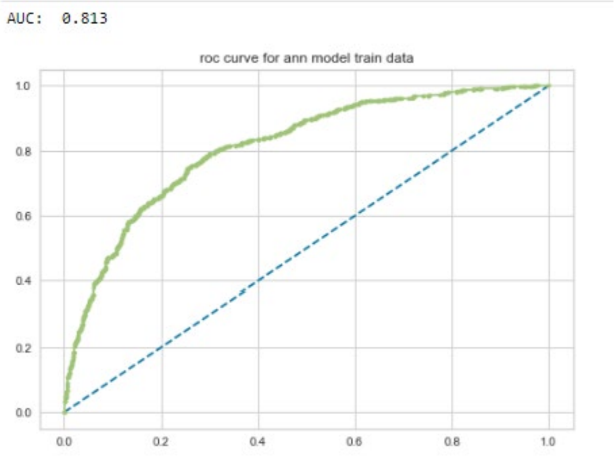
Inferences:

- Train data-Accuracy of train model is 80.0, recall is 0.55, precision is 0.73 and auc is 84.2
- Test data- Accuracy is 76.0, recall is 0.44, precision is 0.73 and auc is 81.5.
- Although the test data is performing close to the train data, we definitely have a room here for improvement in our training data to improve the accuracy further.
- With respect to the Decision Tree (CART) model Random Forest model has performed slightly better in this case.

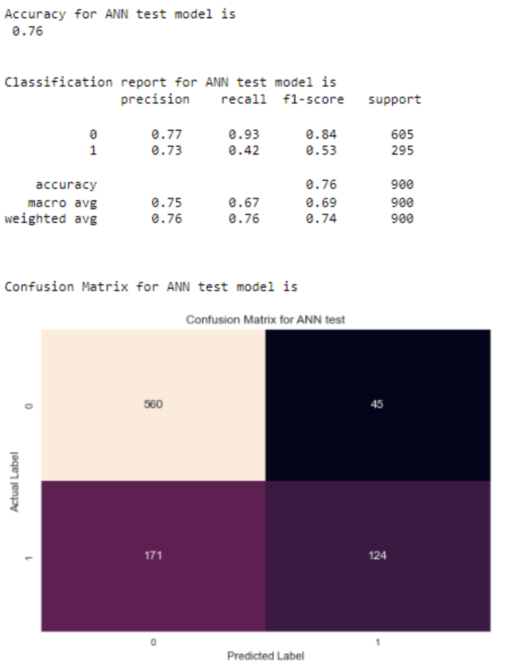
ANN model:



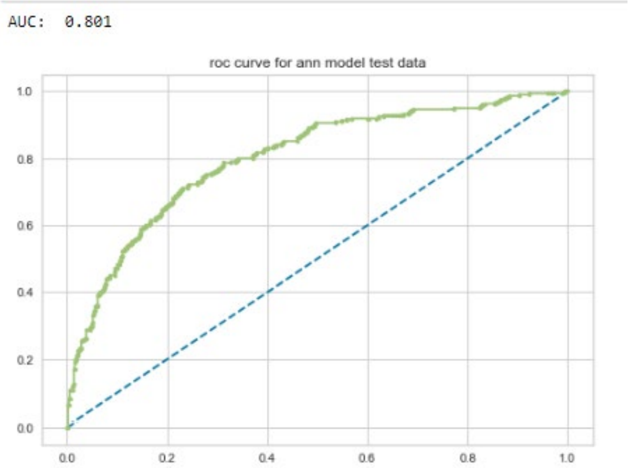
```
Accuracy for ANN training model is
 0.7723809523809524


Classification report for ANN training model is
               precision    recall  f1-score   support

           0       0.80      0.90      0.85      1471
           1       0.67      0.48      0.56       629

    accuracy                           0.77      2100
   macro avg       0.73      0.69      0.70      2100
weighted avg       0.76      0.77      0.76      2100


Confusion Matrix for ANN training model is
```



```
AUC:  0.813
```



```
Accuracy for ANN test model is
 0.76


Classification report for ANN test model is
               precision    recall  f1-score   support

           0       0.77      0.93      0.84       605
           1       0.73      0.42      0.53       295

    accuracy                           0.76       900
   macro avg       0.75      0.67      0.69       900
weighted avg       0.76      0.76      0.74       900


Confusion Matrix for ANN test model is
```



```
AUC:  0.801
```

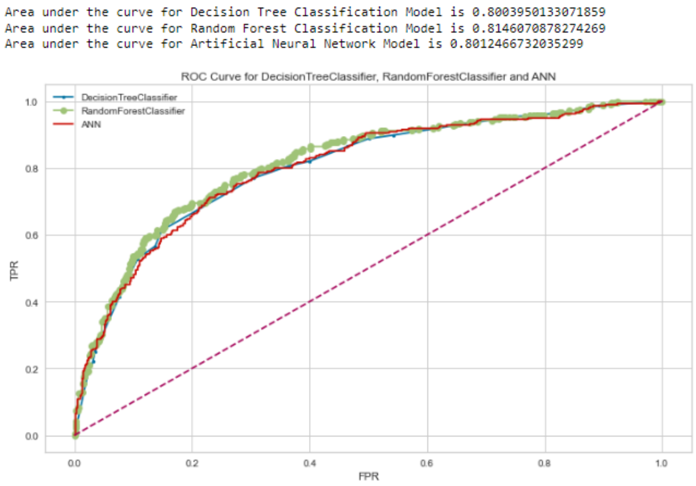
Figure 16

Inferences:

- Train data-Accuracy of train model is 77.0, recall is 0.48, precision is 0.67 and auc is 81.3.4
- Test data-the accuracy is 76.0, recall is 0.67, precision is 0.75 and auc is 80.1.
- Although the test data is performing close to the train data, we definitely have a room here for improvement in our training data to improve the accuracy further.
- With respect to the Decision Tree (CART) model and Random Forest model, we don't see any significant improvement in ANN model.

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

Compare ROC curves and Calculate Area under the curve Decision tree classifier and Random Forest models for test set.



```
Area under the curve for Decision Tree Classification Model is 0.8003950133071859
Area under the curve for Random Forest Classification Model is 0.8146070878274269
Area under the curve for Artificial Neural Network Model is 0.8012466732035299
```

Figure 17

Conclusion:

- RFCL performs better than ANN and Decision Tree on the test set.
- CART and RFCL has the same accuracy score of ~77% while ANN has the least accuracy score of ~76% for test set.
- RFCL and CART model has the highest accuracy score of ~80% while ANN model has the least accuracy score of ~77% for training set.
- CART and ANN have ~80% area under the curve for testing set while RFCL has ~81% area under curve.
- Model can be improved for changing grid parameters for CART and RFCL or by changing hidden layers or tolerance values for ANN model.

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

The business insights and recommendations based on the whole analysis are as follows:

- Agency code: EPX has the maximum count with a reasonably lower claim settlement ratio while C2B and CWT has the maximum number of claims settled. Higher claim settled obviously means revenue loss for the company. JZI need to improve on their business model and try to bring in more customers. EPX business model can be studied in detail and we can try to understand the functionality and work flow of the agency and use it to improve the business models of other agencies.
- Type: Travel agencies have more count than airlines however airlines have a greater number of claims settled. Airline claim settlement policy needs to be reviewed and necessary steps for improvement needs to be taken.
- channel: Online dominates both in count and claim settled. More and more emphasis should be made to improve the marketing strategy and online campaigns to increase the customer base.
- Product name: Customized plan has the maximum count however Gold plan has the maximum number of claims settled with least customer count. These needs to be reviewed and investigated thoroughly.
- Destination: Maximum count is from Asia and Americas has the highest median for claim settled. With respect to count Asia is performing much better than Americas or Europe and we need to review why the median is higher with such low count of customer in this region.

End of problem set2 - CART-RF-ANN.