



11/21/2021


DATA MINING

CART, RANDOM FOREST, ARTIFICIAL
NEURAL NETWORK

INTRODUCTION

This project report is about clustering the data, dividing the customer data into different segments and building your analysis or creating promotional offers for the customers.

Second part is the prediction of future customers about the actions, whether they will classify in class 1 (negative) or classify in class 2 (positive) towards the business.



Contents

Contents	1
PROBLEM 1: CLUSTERING	3
<i>Q1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).</i>	3
<i>Q-1.2. Do you think scaling is necessary for clustering in this case? Justify.</i>	9
<i>Q-1.3 - Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.</i>	10
<i>Q-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.</i>	13
<i>Q-1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.</i>	15
Problem 2: CART-RF-ANN	17
<i>Q-2.1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).</i>	17
<i>Q-2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.</i>	22
<i>Solution:</i>	22
<i>Q 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.</i>	28
<i>Q-2.4-Final Model: Compare all the models and write an inference which model is best/optimized.</i>	34
<i>Q-2.5. Inference: Based on the whole Analysis, what are the business insights and recommendations?</i>	35
TABLES	
Table 1. 1 – TOP 5 ROWS	3
Table 1. 2- DATA DESCRIPTION	4
Table 1. 3 – DATA DESCRIPTION BEFORE SCALING	9
Table 1. 4 – DATA DESCRIPTION AFTER SCALING	9
TABLE 1. 5 – HIERARICAL CLUSTERING	11
TABLE 1. 6- KMeans for different clusters	13
TABLE 1. 7 – FINAL DATASET WITH CUSTOMER SEGMENTATION	15
TABLE 1. 8 – CUSTOMER SEGMENTATION WITH FEATURE	15
Table 2. 1- TOP 5 ROWS	17
Table 2. 2- DATA DESCRIPTION	18
Table 2. 3 – AFTER CHANGING OBJECT TO NUMERICAL	22
Table 2. 4 – MODEL COMPARISON	34

FIGURES

Figure 1. 1 – DISTRIBUTION OF FEATURES..... 5

Figure 1. 2 – OUTLIER’S VISUALIZATION 6

Figure 1. 3 – PAIR PLOT OF FEATURES 7

Figure 1. 4 - CORRELATION PLOT OF FEATURES 8

Figure 1. 5 – BOX PLOT AFTER OULTLIER TREATMENT 10

Figure 1. 6 - DENDROGRAM 10

Figure 1. 7 – ELBOW CURVE..... 13

Figure 2. 1 - DISTRIBUTION PLOT OF FEATURES 19

Figure 2. 2 – BOX PLOT OF FEATURES..... 20

Figure 2. 3 – PAIR PLOT OF FEATURES 20

Figure 2. 4 – CORRELATION OF FEATURES..... 21

Figure 2. 5 – COUNT PLOT OF FEATURES..... 21

Figure 2. 6 – CART FEATURE IMPORTANCE..... 24

Figure 2. 7 – RANDOM FOREST FEATURE IMPORTANCE 26

Figure 2. 8 – FEATURES RELATION FOR BUSINESS..... 35

PROBLEM 1: CLUSTERING

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

DATA DICTIONARY

- 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)

Q1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

SOLUTION:

Top 5 rows of Dataset

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.8752	6.675	3.763	3.252	6.55
15.99	14.89	0.9064	5.363	3.582	3.336	5.144
18.95	16.42	0.8829	6.248	3.755	3.368	6.148
10.83	12.96	0.8099	5.278	2.641	5.182	5.185
17.99	15.86	0.8992	5.89	3.694	2.068	5.837

Table 1. 1 – TOP 5 ROWS

As we can see from the above table the top 5 rows of the data set which describes the credit card usage of different customers.

*As we can check it has
210 rows and 07 features.*

- 1. All the features are
of Float data type.*
- 2. There are no
missing values in
dataset.*
- 3. Total memory
usage by the data
frame is 11.6 KB*

```
<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
```

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210	210	210	210	210	210	210
mean	14.84752	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	0.49148
min	10.59	12.41	0.8081	4.899	2.63	0.7651	4.519
25%	12.27	13.45	0.8569	5.26225	2.944	2.5615	5.045
50%	14.355	14.32	0.87345	5.5235	3.237	3.599	5.223
75%	17.305	15.715	0.887775	5.97975	3.56175	4.76875	5.877
max	21.18	17.25	0.9183	6.675	4.033	8.456	6.55

Table 1. 2- DATA DESCRIPTION

There is plenty of information can be interpreted from the above-described data such as:

1. The minimum spending of customers is around 10000 and goes maximum up to 21000.
2. There is 87 % chance that a customer will pay their full payment amount.
3. The Credit Limit of customers in the data frame ranges between 26000 to 41000.
4. The customers spend good amount of money in a single shopping as their mean is really high around 5500.

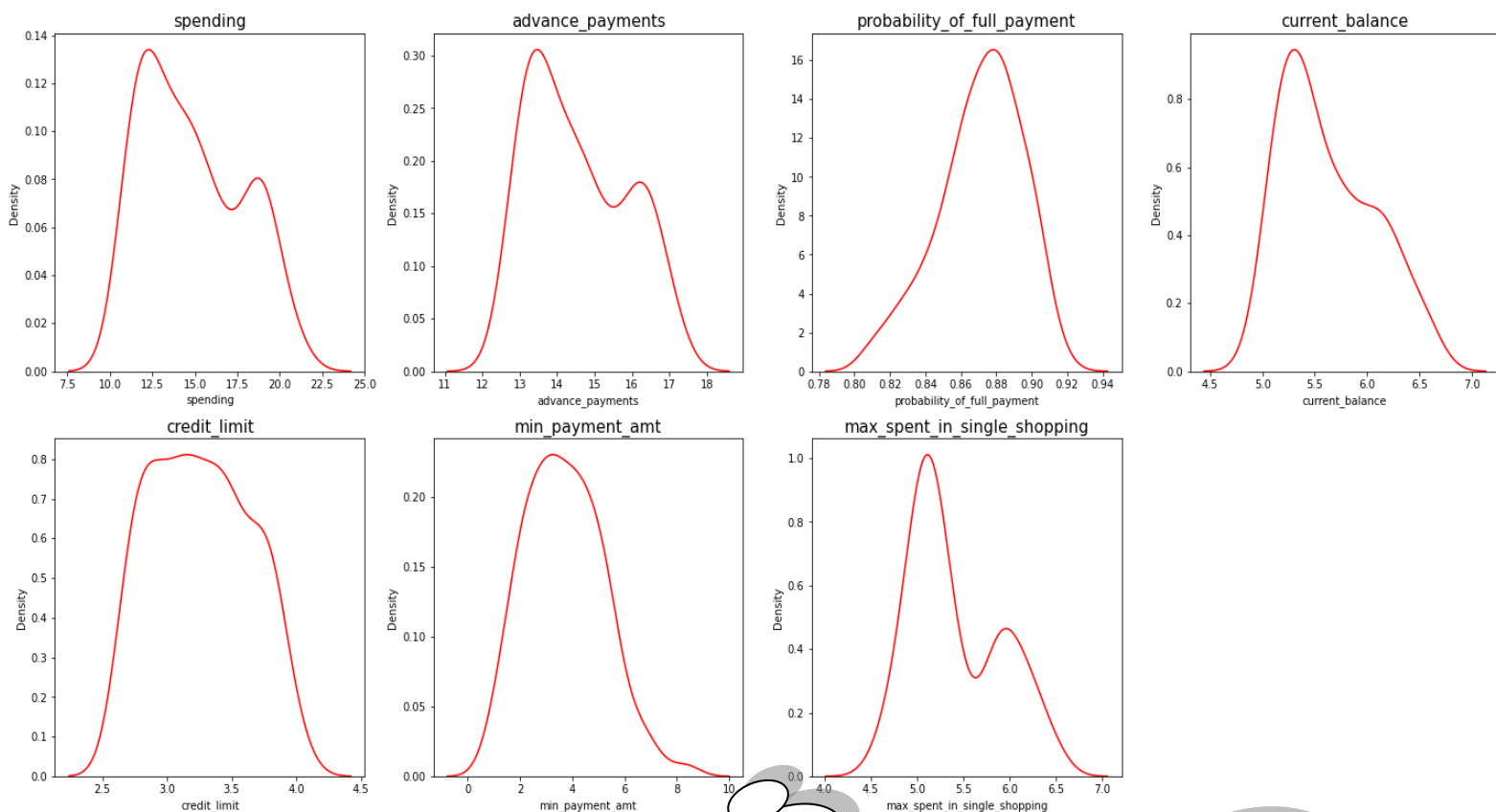


Figure 1.1 – DISTRIBUTION OF FEATURES

As we can check from the histogram plot and skewness stats that,

Spending, Advance_payment and Min_payment_amt is right skewed and has similar skewness values around 0.39

Current_balance and max_spent_in_single_shopping are highly right skewed as they have higher skewness values up to 0.52 and 0.56.

Credit_limit is almost normally distributed as its skewness value is very much close to zero i.e 0.13

Skewness of Spending is 0.3998891917177586

Skewness of advance_payments is 0.3865727731912213

Skewness of probability_of_full_payment is -0.5379537283982823

Skewness of current_balance is 0.5254815601318906

Skewness of credit_limit is 0.1343782451316215

Skewness of min_payment_amt is 0.40166734329025183

Skewness of max_spent_in_single_shopping is

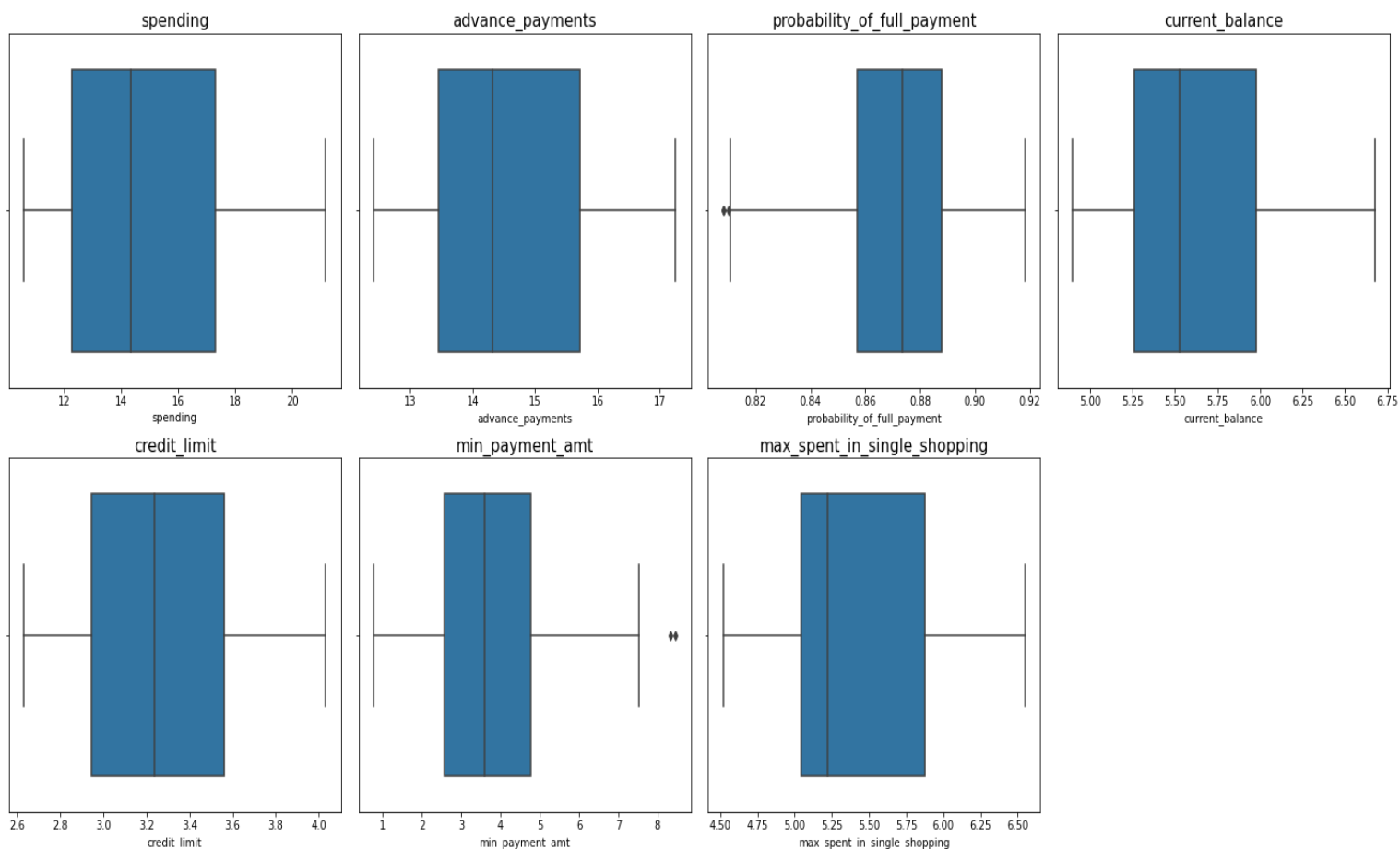


Figure 1. 2 – OUTLIER'S VISUALIZATION

As we can check the box plot of the features there are very few outliers present in the data only in two features that is `probability_of_full_payment` and `min_payment_amt`. In `probability_of_full_payment` maybe outliers represent some customers have very less probability of paying full amount where as in `min_payment_amt` outliers are customers who have paid much more than their minimum amount.

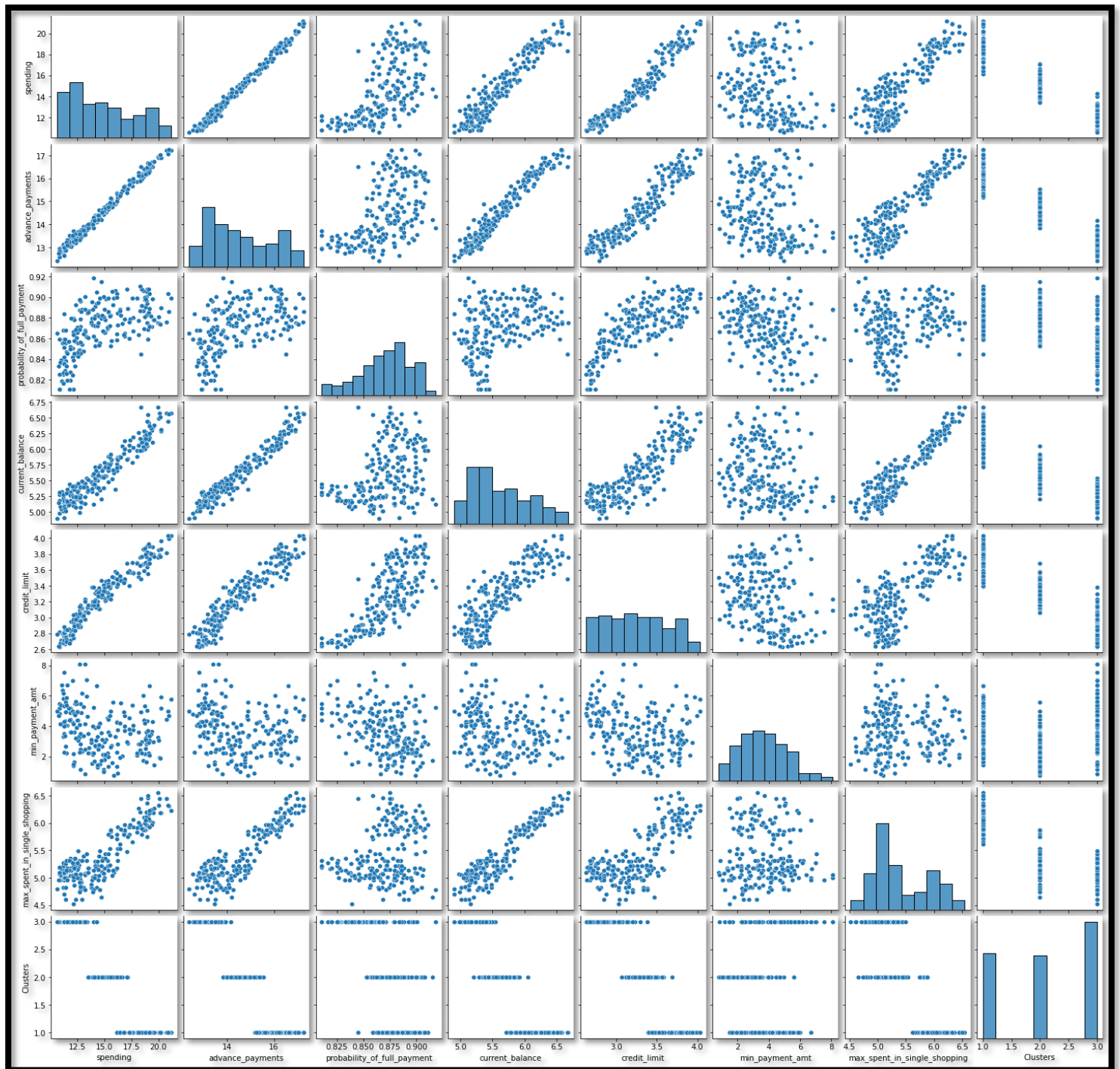


Figure 1. 3 – PAIR PLOT OF FEATURES

As we can check from pair plot there are some correlation between features such as :

1. Spending has positive correlation with advance payments, Credit Limit and Current balance
2. Advance payment has positive Correlation with Credit limit and Current balance.
3. Credit limit and Current balance have correlation.

Pair plot can only show the direction of correlation but not the magnitude.

So,let's check the correlation Heat map to know the magnitude of the correlation.

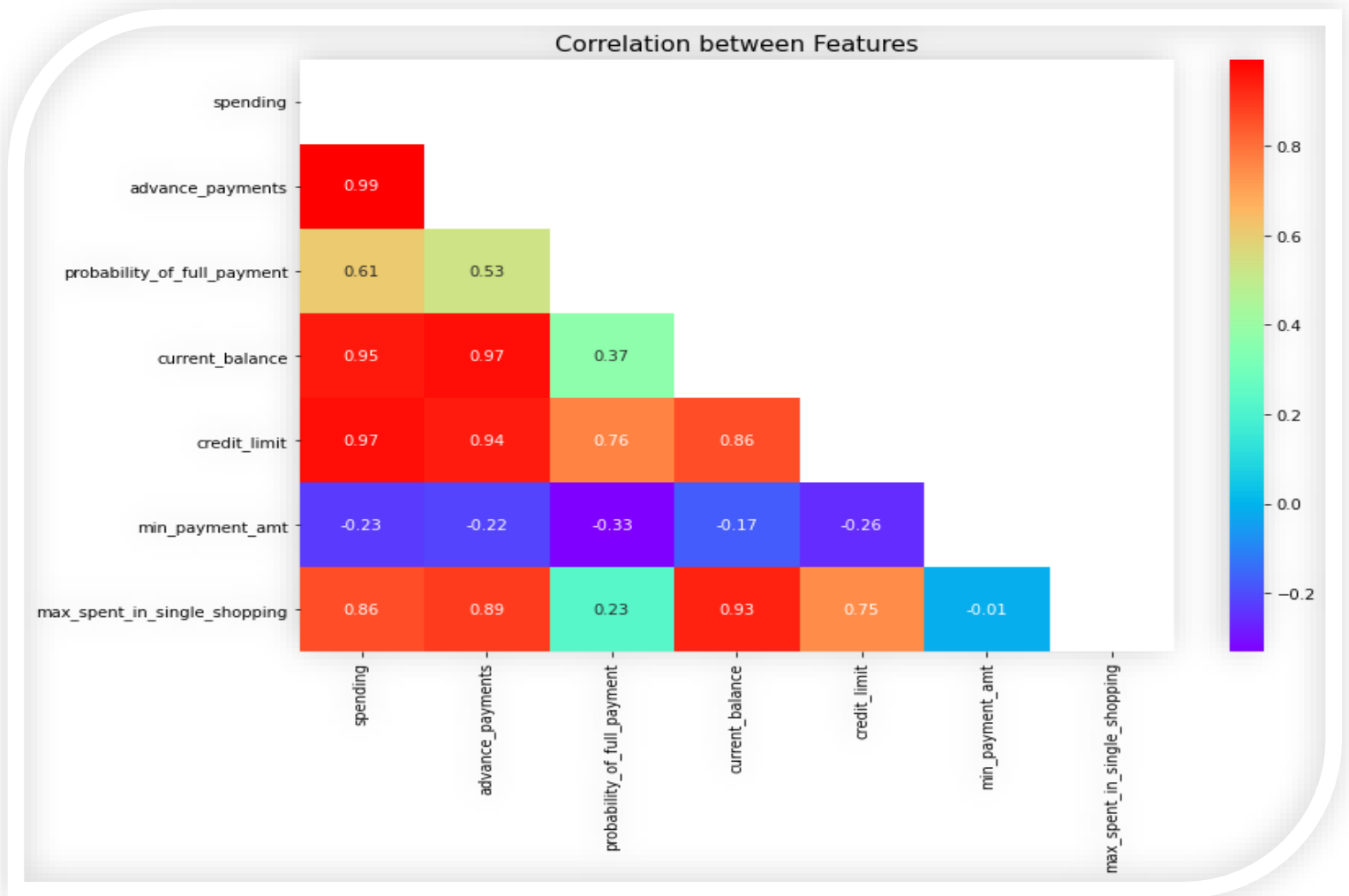


Figure 1. 4 - CORRELATION PLOT OF FEATURES

As we can check from correlation plot

1. There are some features who are highly correlated with each other's such as:
 spending with advance_paymnets,
 spending-current_balance,
 spending-credit_limit
 which represents spending has very high impact on these features like if spending will increase current balance increase and credit limit will decrease.
2. advance_payment is also highly correlated with credit_limit and current_limit as if a customer is doing advance payments, then its current balance will reduce and credit limit will increase.
3. Current balance is also highly related with credit limit and max spent in single shopping.

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

Solution:

Scaling is necessary because of different ranges of features:

DESCRIPTION OF THE DATA SET							
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210	210	210	210	210	210	210
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	0.49148
min	10.59	12.41	0.8081	4.899	2.63	0.7651	4.519
25%	12.27	13.45	0.8569	5.26225	2.944	2.5615	5.045
50%	14.355	14.32	0.87345	5.5235	3.237	3.599	5.223
75%	17.305	15.715	0.887775	5.97975	3.56175	4.76875	5.877
max	21.18	17.25	0.9183	6.675	4.033	8.456	6.55

Table 1. 3 – DATA DESCRIPTION BEFORE SCALING

As we can check from the above description of data frame needs to scaled because

- 1. All the features have different value ranges like some of them are in 1000s some of them in 100s and even one of them in 10000s.
- 2. From detail description of data, we can check that their means are very distinctive from each other it ranges from 14.84 to 0.87.
- 3. As like mean standard deviation of each feature is also varying.

DESCRIPTION OF DATA AFTER SCALING							
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	2.10E+02	2.10E+02	2.10E+02	2.10E+02	2.10E+02	2.10E+02	2.10E+02
mean	9.15E-16	1.10E-16	1.24E-15	-1.09E-16	-2.99E-16	5.30E-16	-1.94E-15
std	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
min	-1.47E+00	-1.65E+00	-2.67E+00	-1.65E+00	-1.67E+00	-1.96E+00	-1.81E+00
25%	-8.88E-01	-8.51E-01	-5.98E-01	-8.29E-01	-8.35E-01	-7.59E-01	-7.40E-01
50%	-1.70E-01	-1.84E-01	1.04E-01	-2.38E-01	-5.73E-02	-6.75E-02	-3.77E-01
75%	8.47E-01	8.87E-01	7.12E-01	7.95E-01	8.04E-01	7.12E-01	9.56E-01
max	2.18E+00	2.07E+00	2.01E+00	2.37E+00	2.06E+00	3.17E+00	2.33E+00

Table 1. 4 – DATA DESCRIPTION AFTER SCALING

For scaling the data, we used Standard Scaler (Z Score) method and after scaling the data we can compare both data description

- 1. All the features are now in the same range.
- 2. All feature’s mean is nearly 0 and standard deviation is around 1.

STANDARD SCALER (Z SCORE)
Standard Scaler is method of scaling the data in which it makes the mean of every feature to zero and then calculate the distance of each data points from its mean in terms standard deviation.

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

Solution:

Before applying the Hierarchical Clustering, we need to treat the outliers in the data set as it outliers can affect the performance of clustering.

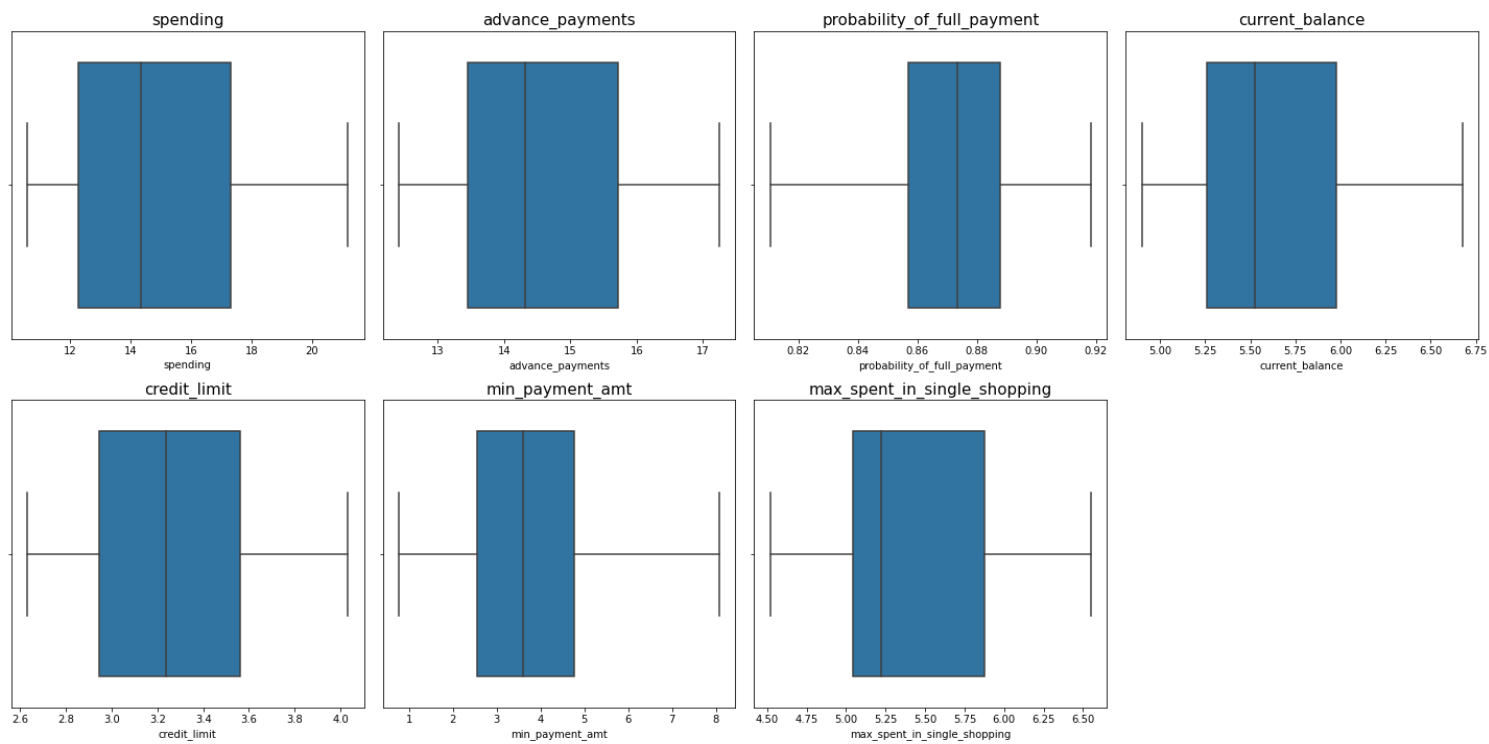


Figure 1. 5 – BOX PLOT AFTER OULTIER TREATMENT

As we can check in the above graphs that there no more outlier's presents in the data set.

Now it's ready to perform hierarchical clustering.

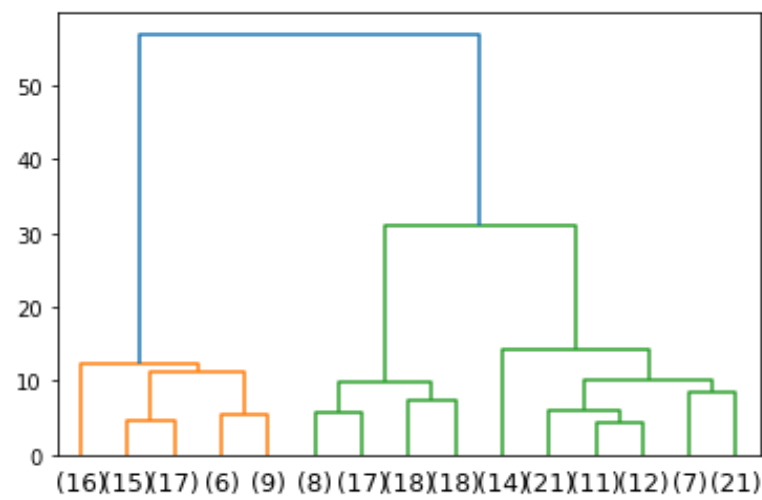


Figure 1. 6 - DENDROGRAM

DENDROGRAM

This is the graphical representation of cluster formation of last 15 clusters.

The numbers represent that how many data points are in one cluster. At the bottom of graph there are 15 clusters with different number of data points in it, as it goes further to form one cluster at the end.

Hierarchical Clustering works bottom to top as it starts with each data point as one cluster then joining it with other clusters on the method until all data points come in one cluster.

For clustering we chose **WARD** as a linkage method for clustering as it was resulting in better distribution of data points into 3 clusters as compared to other methods. So finally, we are choosing 3 clusters to represent the customer data.

CLUSTER OF DATA								
Index	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	6.55	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	3
4	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1
...
205	13.89	14.02	0.888	5.439	3.199	3.986	4.738	2
206	16.77	15.62	0.8638	5.927	3.438	4.92	5.795	1
207	14.03	14.16	0.8796	5.438	3.201	1.717	5.001	2
208	16.12	15	0.9	5.709	3.485	2.27	5.443	2
209	15.57	15.15	0.8527	5.92	3.231	2.64	5.879	2

TABLE 1. 5 – HIERARCHICAL CLUSTERING

We have chosen 3 Clusters for the business problem as it represents the customer segmentation better than any other and easier to study and make business decisions.

Only 2 clusters don't make any business sense as it only represents high spending customers and low spending customers and it will not do any justification for the business problem.

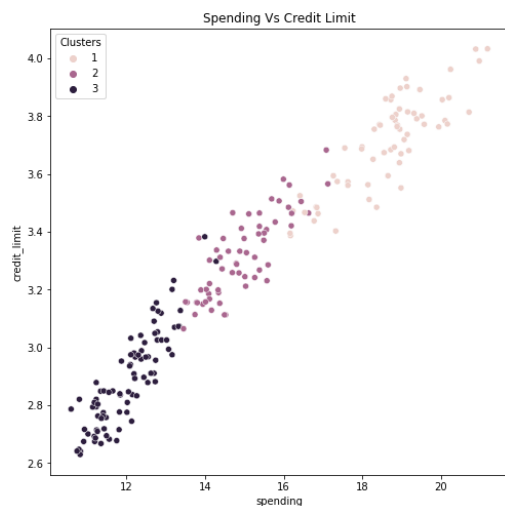
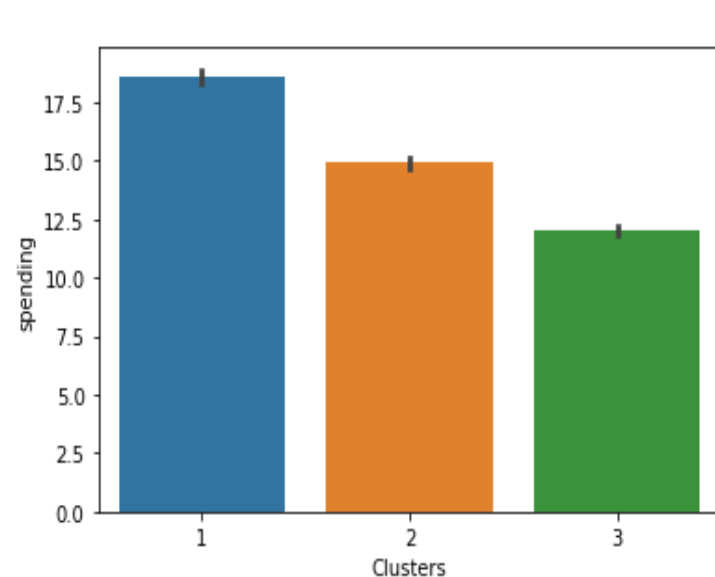
If we choose more than 3 clusters, as per shown in dendrogram it will create 5 clusters and that will make the result more complex and difficult to study for making business decision.

This column represents that which data point belongs to which cluster in total three clusters.

WARD METHOD

Ward's method starts with n clusters, each containing a single object. These n clusters are combined to make one cluster containing all objects. At each step, the process makes a new cluster that minimizes variance, measured by an index called E (also called the sum of squares

CUSTOMER SEGMENTATION								
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
Clusters								
1	18.615873	16.259524	0.883937	6.194603	3.708143	3.659413	6.060952	63
2	14.920164	14.579344	0.881487	5.608033	3.314852	2.684805	5.221951	61
3	12.035465	13.299535	0.854146	5.228395	2.889395	4.44319	5.061814	86



As we infer from the plots:

1. Cluster 1 customers have higher spending per month and their Credit limit and Current balance is also higher.
2. Cluster 2 have medium spending lower than cluster 1 but higher than cluster 3 also same with their Credit Limit and Current balance.
3. Cluster 3 has the lowest spending and low Credit limit, Current balance amongst all three clusters.



Freq column represents the customer segmentation in each Cluster and each features has their average value for each Cluster.

Cluster 1- High Spenders – As they have high spending average almost 50% of their Credit limit.

Cluster 2- Medium Spenders – As they have medium Spending average not too high or too low (45 % of credit limit.)

Cluster 3- Low Spenders – They have the lowest average in spending. (41 % of Credit Limit)



Q-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:

KMeans is another method for clustering. In this method first we have to provide the values of clusters and then choose the optimum number of clusters by checking their values and scores.

Applying KMeans clustering with 1-10 numbers of clusters. And Inertia,silhouette scores for each number of clustering and visualizing it with elbow curve.

KMEANS CLUSTERING		
Clusters	Kmeans inertia	Silhouette score
1	1470	
2	659.147401	0.465601004
3	430.2984818	0.400805922
4	371.2217639	0.329437337
5	326.8846408	0.286487372
6	290.1513312	0.284655034
7	263.0291033	
8	242.8107073	
9	221.487597	
10	206.3290465	

TABLE 1. 6- KMeans for different clusters

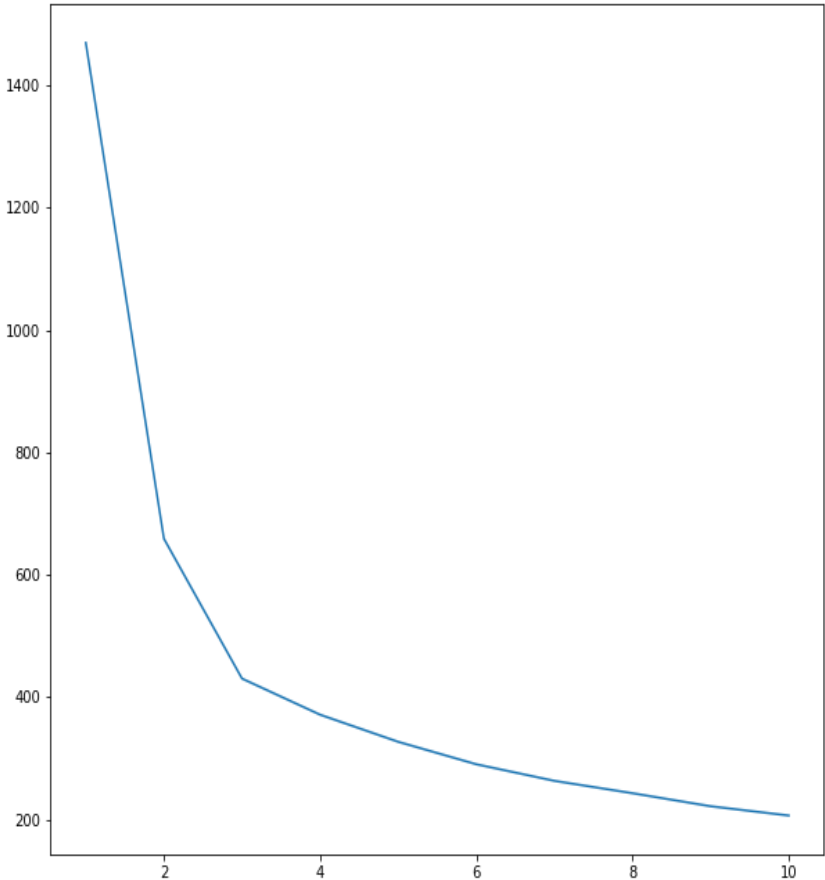
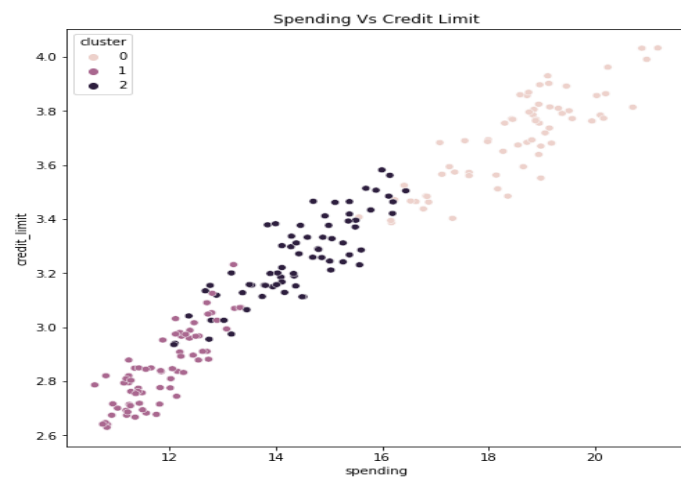
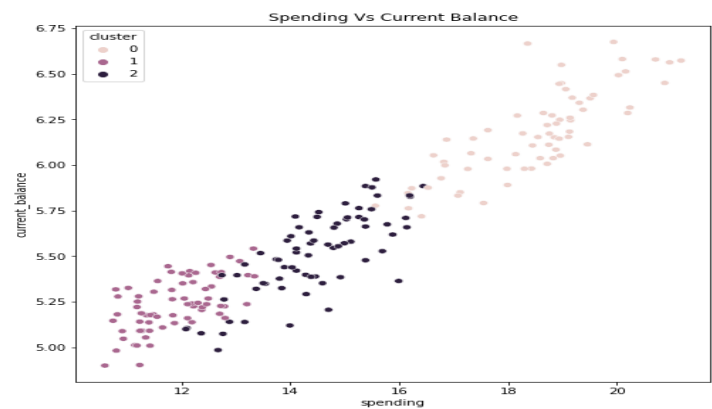
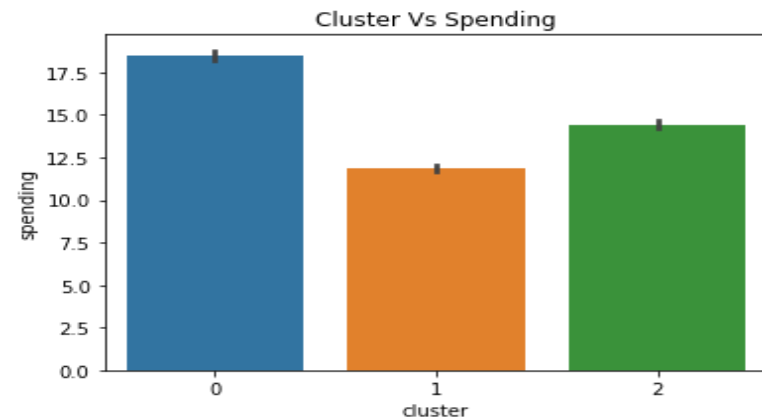


Figure 1. 7 – ELBOW CURVE

The above table comprises the different inertia and silhouette scores for different cluster numbers from 1-10, the drop of inertia values is higher from 1 to 2 and then from 2 to 3. But after 3 the drop in the inertia values is much lower as compare to above two. In the elbow curve also we can observe the same, **so the optimum number of Clusters should be 3 using KMeans clustering. The silhouette score for 3 clusters is 0.40 which better than other clusters.** Choosing only 2 clusters will not make much business sense it will segment the customers in only high and low spending customers and deriving business insights will not justify the business problem.

CUSTOMER SEGMENTATION USING KMEANS								
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
cluster								
0	18.495373	16.203433	0.88421	6.175687	3.697537	3.632373	6.041701	67
1	11.856944	13.247778	0.84833	5.23175	2.849542	4.733892	5.101722	72
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	71



As we infer from theplots:

1. Cluster 0, customers have higher spending per month and their Credit limit and Current balance is also higher.
2. Cluster 2, have medium spending lower than cluster 1 but higher than cluster 3 also same with their Credit Limit and Current balance.
3. Cluster 1, has the lowest spending and low Credit limit, Current balance amongst all three clusters.



Freq column represents the customer segmentation in each Cluster and each feature has their average value for each Cluster.

Cluster 0- High Spending Customers– As they have high spending average almost 50% of their Credit limit.

Cluster 2– Medium Spending Customers – As they have medium Spending average not too high or too low (44 % of credit limit.) and their credit limit is also near to high spending customer but spending is less than cluster 0 customers.

Cluster 1– Low Spending Customers – They have the lowest average in spending.(41 % of Credit Limit). They have the lowest credit card limit and spending is also lower than other cluster customers,



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

Solution:

After clustering the data frame using both hierarchal and KMeans clustering, we checked that KMeans customer segmentation in not properly done as well as their silhouette score is also low for 3 clusters (0.40). So we are moving forward with hierarchal clustering and adding the cluster to the original data frame and performing Cluster profiling.

DATASET WITH CLUSTERS								
Index	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	6.55	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	3
4	17.99	15.86	0.8992	5.89	3.694	2.068	5.837	1
...
205	13.89	14.02	0.888	5.439	3.199	3.986	4.738	2
206	16.77	15.62	0.8638	5.927	3.438	4.92	5.795	1
207	14.03	14.16	0.8796	5.438	3.201	1.717	5.001	2
208	16.12	15	0.9	5.709	3.485	2.27	5.443	2
209	15.57	15.15	0.8527	5.92	3.231	2.64	5.879	2

TABLE 1. 7 – FINAL DATASET WITH CUSTOMER SEGMENTATION

As we can check Customer segmentation has properly completed and finally has been added the Dataset.

Now let’s check each feature mean with respect to the clustering.

CUSTOMER SEGMENTATION								
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	Freq
Clusters								
1	18.615873	16.259524	0.883937	6.194603	3.708143	3.659413	6.060952	63
2	14.920164	14.579344	0.881487	5.608033	3.314852	2.684805	5.221951	61
3	12.035465	13.299535	0.854146	5.228395	2.889395	4.44319	5.061814	86

TABLE 1. 8 – CUSTOMER SEGMENTATION WITH FEATURE

Cluster Profiling

Cluster-1- High Spending Customer.

I have profiled the Cluster 1 as high spending Customer for following reason:

- 1. They have highest spending in a month amongst all the clusters.
- 2. The Monthly Spending is almost 50 % of their credit limit.
- 3. They have the highest amount for maximum spent in single shopping.

These customers are high spending customers but they also have the highest probability of paying the full amount at once, it can be good for the bank but also a bit riskier. So we have to approach them very carefully and create promotional offer specifically targeting their needs depending on their transaction history.

Promotional Strategies:

1. We can offer them up gradation of their card with added benefits.
2. Promotional offers on the frequent purchasing items.

Cluster -2- Medium Spending Customers.

I have profiled the Cluster 2 as medium spending customer because of the following reason:

1. These customers have medium average monthly spending, Lower than cluster 1 and higher than cluster 3.
2. Their credit limit is also as same as spending and their monthly spending is around 45 % of credit limit.

These Customers have a good potential to be in the higher segment as they have good credit limit almost near to high segment customers, it's just their spending is less. They also have the high probability of paying the full amount, So need to address their needs and create different promotional offers depending on their transaction history to increase their spending.

Promotional Strategies:

1. Cash backs offers on minimum spent amount.
2. Increasing the Credit limit of the customers.

Cluster -3 – Low Spending Customers.

I have profiled the Cluster 3 as low spending customers for the following reasons:

1. These customers have lowest monthly spending as compare to other clusters almost 41 % of Credit Limit.
2. Their Credit limit is also lesser than the other customer segment and still they have low monthly spending.

These customers are low spending customers as well as have lowest probability of full payment amongst the entire three segments. So we need to encourage them to spent more and find the customers who have high current balance and lowest probability to full payment and offer them different payment options.

Promotional Strategies:

1. Promotional and discount Vouchers for different E-commerce platform.
2. Easy or interest free EMI options for big amount purchase.
3. EMI options for reducing the balance amount.

Problem 2: CART-RF-ANN

An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

DATA DICTIONARY

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)

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

Solution:

TOP 5 ROWS OF DATASET										
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0	Online	34	20	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0	Online	4	26	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.3	Online	53	18	Bronze Plan	ASIA

Table 2. 1- TOP 5 ROWS

As we can check from the above table the top 5 rows of the dataset which provides the information of customers Tour details like destination and duration, it also provides the insurance plan the purchased through which type like Agency or Airlines and how much sales made through the selling of insurance.

As we can check from the data information,

1. It has 3000 rows and 10 columns.
2. There are 02-Float64, 02-Int64 and 06-Object types of data present in the data frame.
3. There are no missing values present in the data.
4. As all the feature data type are in order and there is no missing value, there is no anomalies present in the data frame.

```
<class 'pandas.core.frame.DataFrame'>
```

RangeIndex: 3000 entries, 0 to 2999

Data columns (total 10 columns):

#	Column	Non-NullCount	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

DATA DESCRIPTION

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000	3000	3000	3000	3000	3000	3000	3000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	2954	NaN	NaN	1136	2465
mean	38.091	NaN	NaN	NaN	14.529203	NaN	70.00133	60.25	NaN	NaN
std	10.463518	NaN	NaN	NaN	25.481455	NaN	134.0533	70.734	NaN	NaN
min	8	NaN	NaN	NaN	0	NaN	-1	0	NaN	NaN
25%	32	NaN	NaN	NaN	0	NaN	11	20	NaN	NaN
50%	36	NaN	NaN	NaN	4.63	NaN	26.5	33	NaN	NaN
75%	42	NaN	NaN	NaN	17.235	NaN	63	69	NaN	NaN
max	84	NaN	NaN	NaN	210.21	NaN	4580	539	NaN	NaN

Table 2. 2- DATA DESCRIPTION

From the above data description:

1. The people traveled and bought travel insurance has a age range between 08 to 84 years.
2. Most of the customers around 61 % that took insurance are through travel agency.
3. Most of the customer has not claimed any insurance.
4. The average duration of customer tours is around 70 days.
5. The average sales of insurance are around 6000.

There are 139 duplicate rows in the data frame, as this is an insurance data and same packages could have sold to multiple customers. So this data can have crucial information and removing it can impact the model performance. So keeping the duplicated rows for further analysis.

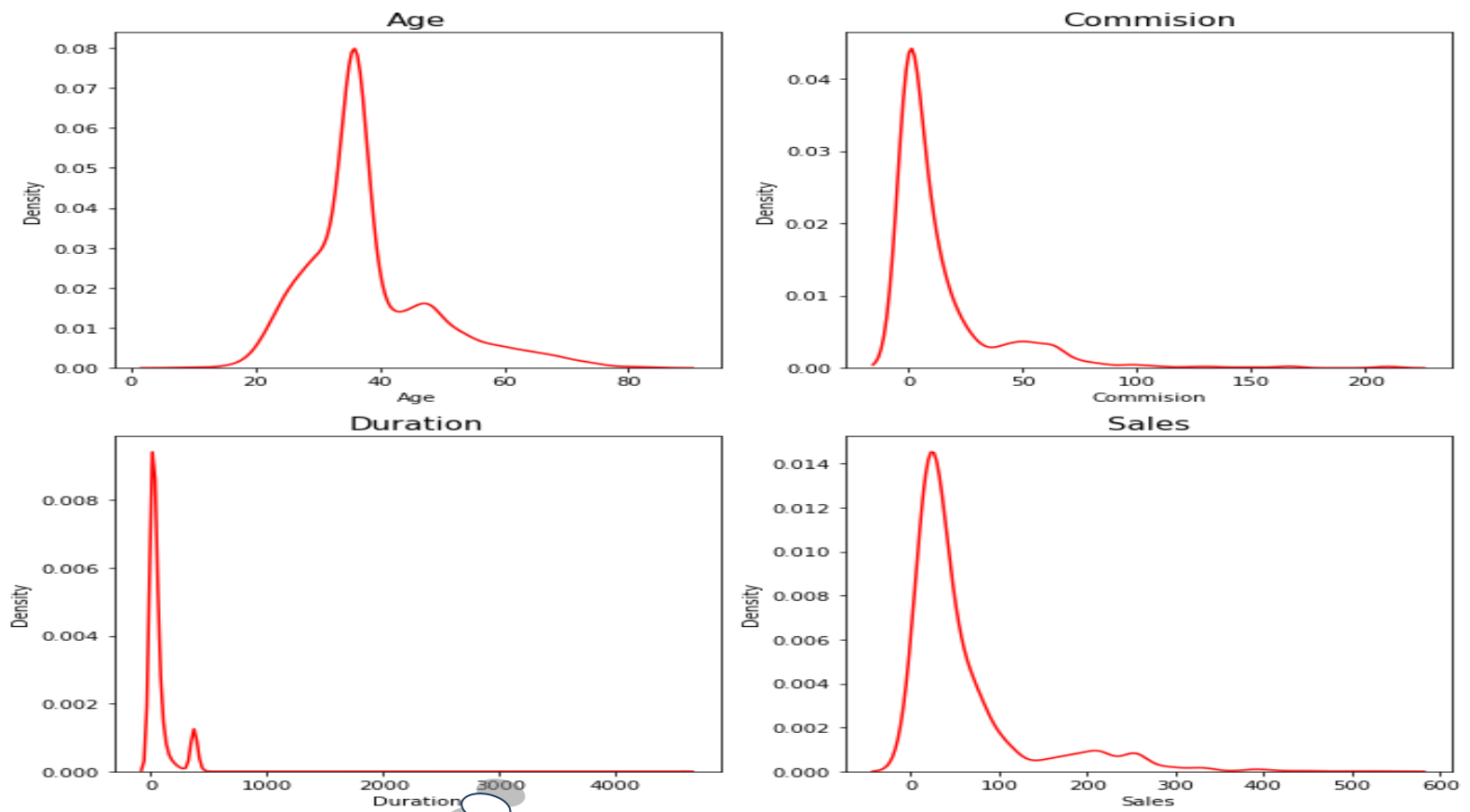


Figure 2. 1 - DISTRIBUTION PLOT OF FEATURES

Skewness of Age is 1.149712770495169

Skewness of Commission is
3.148857772356885

Skewness of Duration is
13.784681027519602

Skewness of Sales is
2.381148461687274

As we can check from the plot and information that all the numerical columns are highly right skewed.

Duration has a very high skewness values as compare to others.

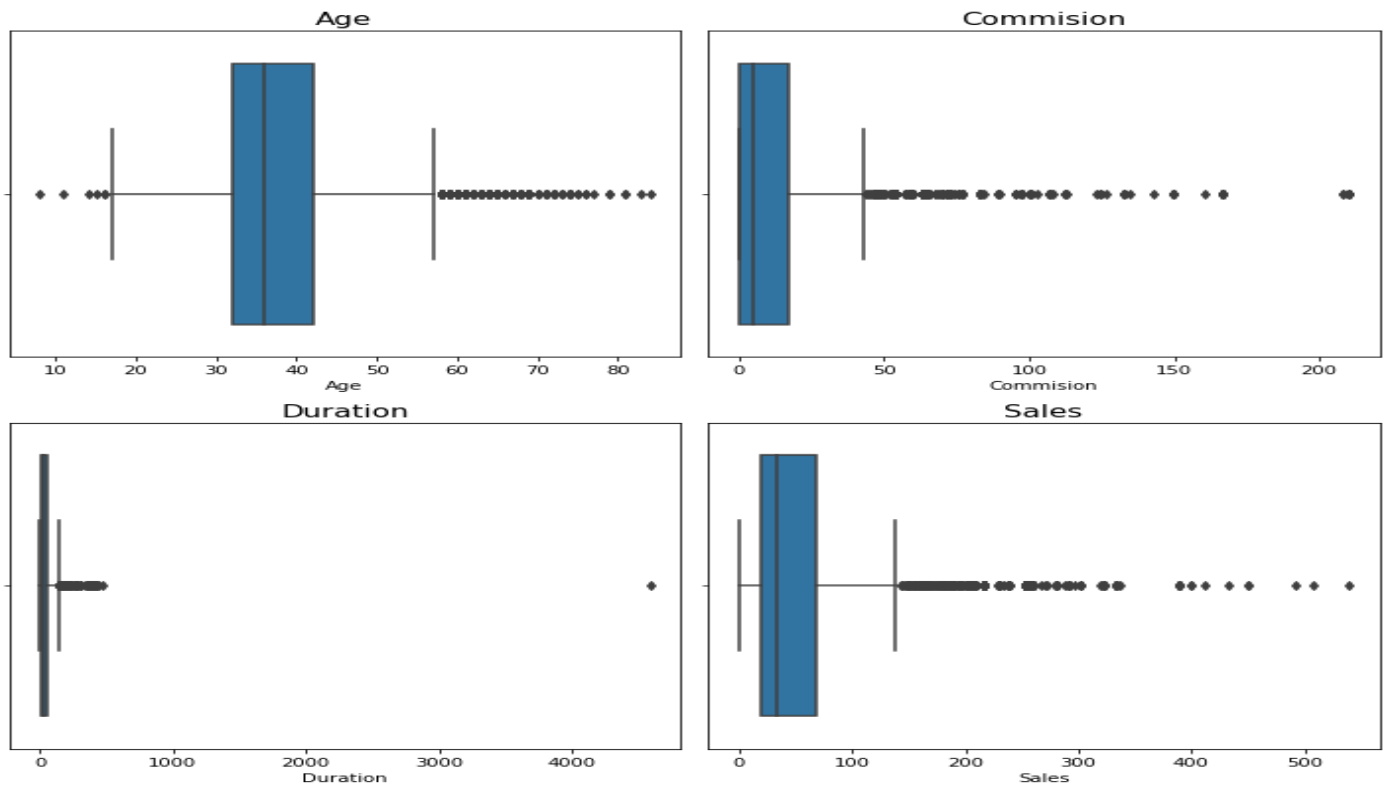


Figure 2. 2 – BOX PLOT OF FEATURES

Box plot of the numerical columns also shows that Each and every numerical feature has outliers in it . The number of outliers is also large so treating it might impact the model performance.

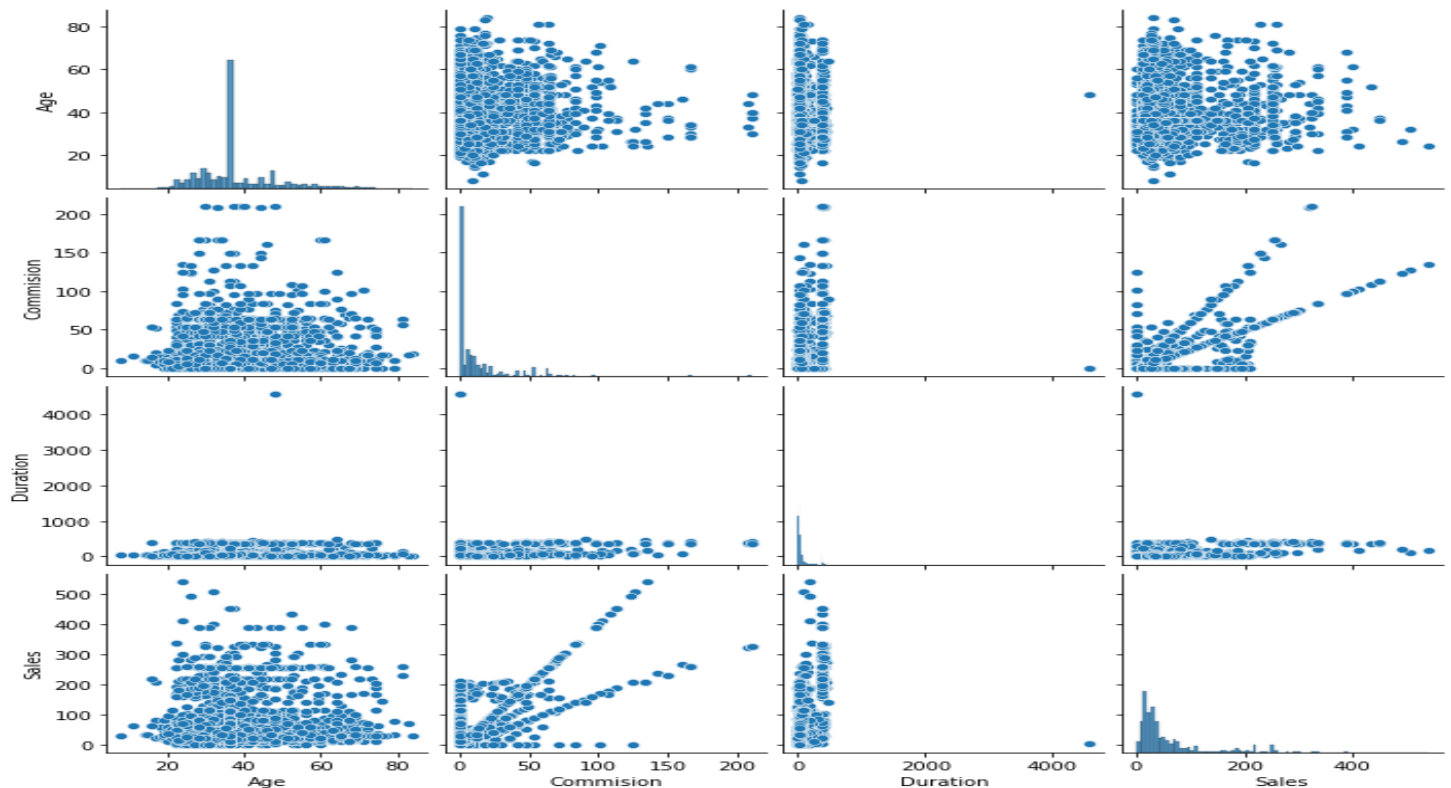
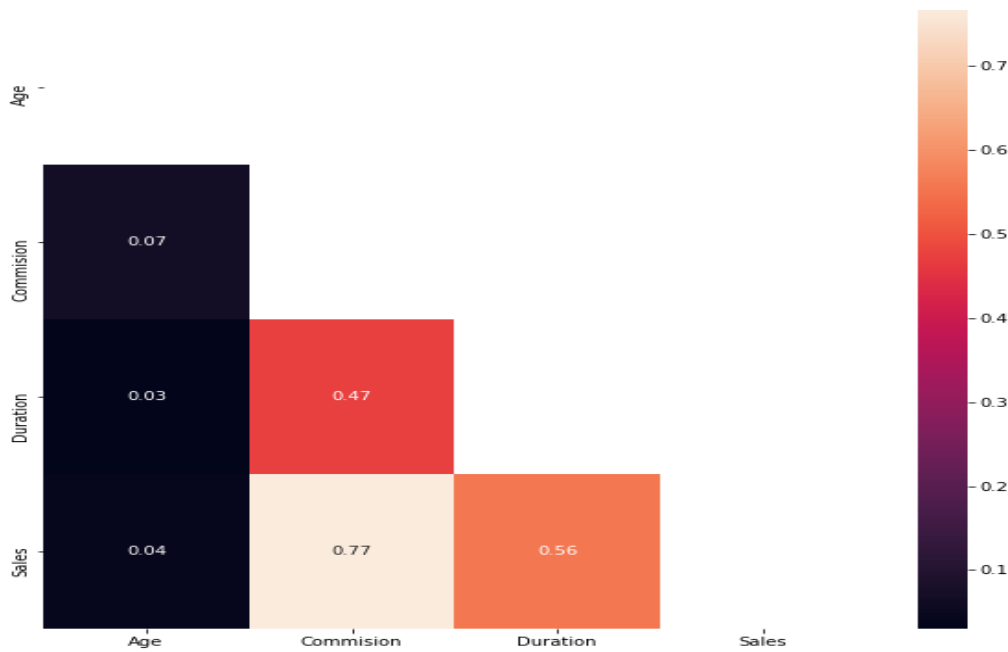


Figure 2. 3 – PAIR PLOT OF FEATURES



As we can check from the above correlation plot that there are not much correlation present in the data set, Only Sales to commission and sales to duration are positively correlated.

Higher the sales, higher the commission and longer the duration of the tour higher the sales.

Figure 2. 4 – CORRELATION OF FEATURES

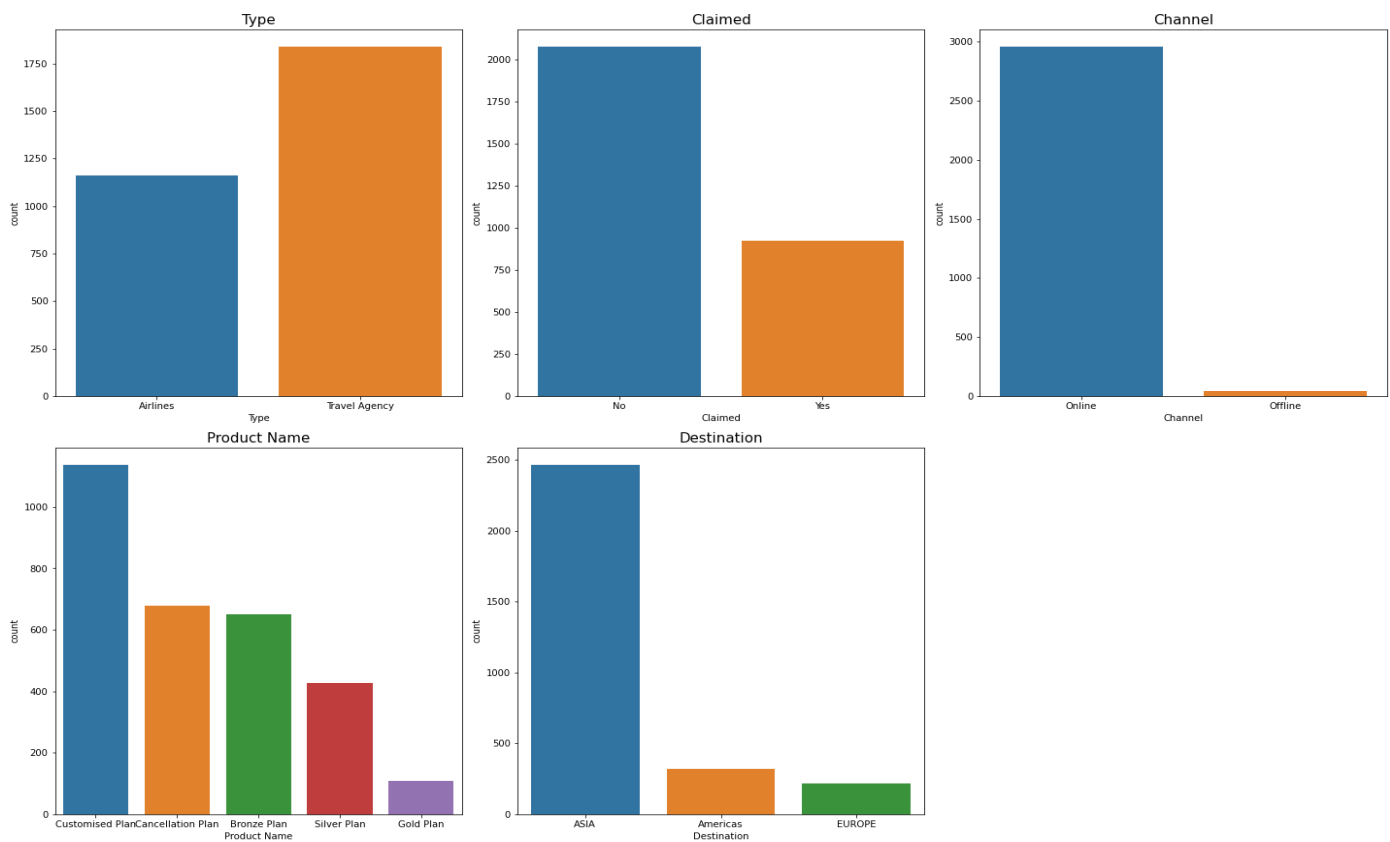


Figure 2. 5 – COUNT PLOT OF FEATURES

From the above we can easily interpret that:

1. Travel agency is preferred mode of travel insurance rather than airline.
2. Online platforms are most used for purchasing travel insurance for the tour.
3. Most Customer prefer customized plan for their insurance while traveling and after that cancellation plan.
4. Asia is the most preferred or liked destination for travelling.

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

Solution:

TOP 5 ROWS OF DATASET										
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0	Online	34	20	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0	Online	4	26	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.3	Online	53	18	Bronze Plan	ASIA

As there are object data type present the data frame we need to convert the object data frame into numerical as models only take numerical value.

CONVERTING OBJECT DATATYPE TO NUMERICAL										
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	0	0	0.7	1	7	2.51	2	0
1	36	EPX	1	0	0	1	34	20	2	0
2	39	CWT	1	0	5.94	1	3	9.9	2	1
3	36	EPX	1	0	0	1	4	26	1	0
4	33	JZI	0	0	6.3	1	53	18	0	0

Table 2. 3 – AFTER CHANGING OBJECT TO NUMERICAL

We used the categorical codes to convert the object data type to numerical. Agency Code is the code provided to different insurance packages; it doesn't have any use for further analysis of the data. So, dropping the column from the data frame.

For splitting the data, we are using the method of **train and test split method**.

Before Splitting the data frame first, we need to check whether the data is balance or not,

0 - 0.692 or 69.2 %
1 - 0.308 or 30.8 %

We can check the data is balanced, we can proceed further into splitting the data into test train, For that we are choosing to split the data into 70 & 30 % , 70 % data will used for training and 30 % data will be used for testing the data.

After performing train/test split

X train = (2100, 8)
X test = (900, 8)
y train = (2100)
y test = (900)

X train– Contains 70 % features for training.
X test – Contains 30% features for training.
Y train –Contains 70% target variable for training.
Y test - Contains 30% target variable for testing the data.

TRAIN/TEST SPLIT

The data we use is usually split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

Balanced Data frame

Before splitting the data, we have to check the data frame is balanced or not. For that we have to check the target variable as it is in 0 & 1, whether these two are balanced divided or not which is 70 to 30 ratio.

CREATING DECISION TREE MODEL.

CART METHOD

A Classification and Regression Tree (CART), is a **predictive model**, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

Creating model using cart technique.

Parameters used to create model

1. Criterion – Gini
2. Maximum depth = 10
3. Minimum sample leaf = 30
4. Minimum sample split = 100

We have chosen these parameters using Grid search CV method which provides best possible outcome using given parameters.

REPORT FOR TRAIN DATA

	precision	recall	f1-score	support
0	0.84	0.87	0.85	1471
1	0.67	0.62	0.64	629
accuracy			0.79	2100
macro avg	0.75	0.74	0.75	2100
weighted avg	0.79	0.79	0.79	2100

REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.80	0.88	0.83	605
1	0.68	0.54	0.60	295
accuracy			0.77	900
macro avg	0.74	0.71	0.72	900
weighted avg	0.76	0.77	0.76	900

GINI

Gini index is a CART algorithm which measures a distribution among affection of specific-field with the result of instance. It means, it can measure how much every mentioned specification is affecting directly in the resultant case.

MAXIMUM DEPTH

It is the maximum depth of any node of the final tree, counted from the root node.

MINIMUM SAMPLE SPLIT

Minimum samples leaf specifies the minimum number of samples required to be at a leaf node after splitting the node. If the requirement does not match then it terminates the node in the previous node.

MINIMUM SAMPLE SPLIT

Minimum samples split specifies the minimum number of samples required to split an internal node, If the requirement does not match it will not split further in leaf nodes and terminate the node there.

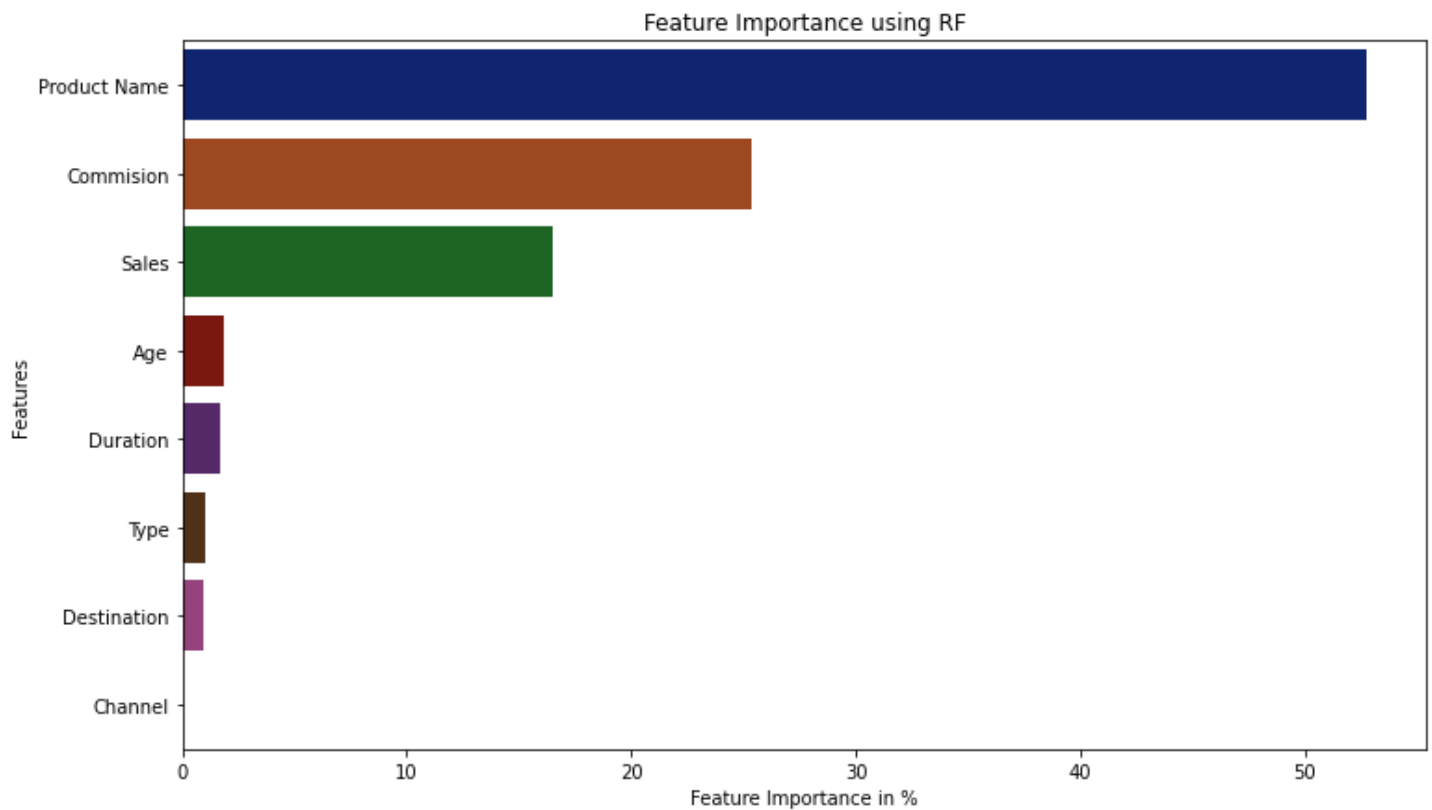


Figure 2. 6 – CART FEATURE IMPORTANCE

This graph represents the features importance to build the decision tree model. Product name plays the most important roll more than 50 % to build this model followed by commission with around 28 % then sales around 18 % then rest 4 % is contributed by rest of features. Channel is the least important feature to build this model.

This importance is decided by using the Gini Index, which ever feature has the highest gini index value plays the most important roll to build the model.

Creating Random Forest Model

Random forest is a supervised learning algorithm which is used for both classification as well as regression. But however, it is mainly used for classification problems. As we know that a forest is made up of trees and more trees means more robust forest. Similarly, random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting. It is an ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result.

Creating the model

Hyper - Parameters used for tuning the model are:

1. Maximum depth = 10
2. Maximum features = 5
3. Minimum sample leaf = 30
4. Minimum sample split = 100
5. Number of estimators = 501

We have chosen these parameters using Grid search CV method which provides best possible outcome using given parameters.

REPORT FOR TRAIN DATA

	precision	recall	f1-score	support
0	0.81	0.91	0.86	1471
1	0.71	0.50	0.59	629
accuracy			0.79	2100
macro avg	0.76	0.71	0.72	2100
weighted avg	0.78	0.79	0.78	2100

REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.76	0.92	0.83	605
1	0.71	0.39	0.51	295
accuracy			0.75	900
macro avg	0.73	0.66	0.67	900
weighted avg	0.74	0.75	0.73	900

PARAMETER DESCRIPTION

MAXIMUM FEATURES

These are the maximum number of features Random Forest is allowed to try in individual tree.

NUMBER OF ESTIMATORS

This is the number of trees you want to build before taking the maximum voting or averages of predictions. Higher number of trees gives you better performance but makes your code slower.

Maximum depth, Minimum sample leaf and Minimum sample split are the same as Decision tree model.

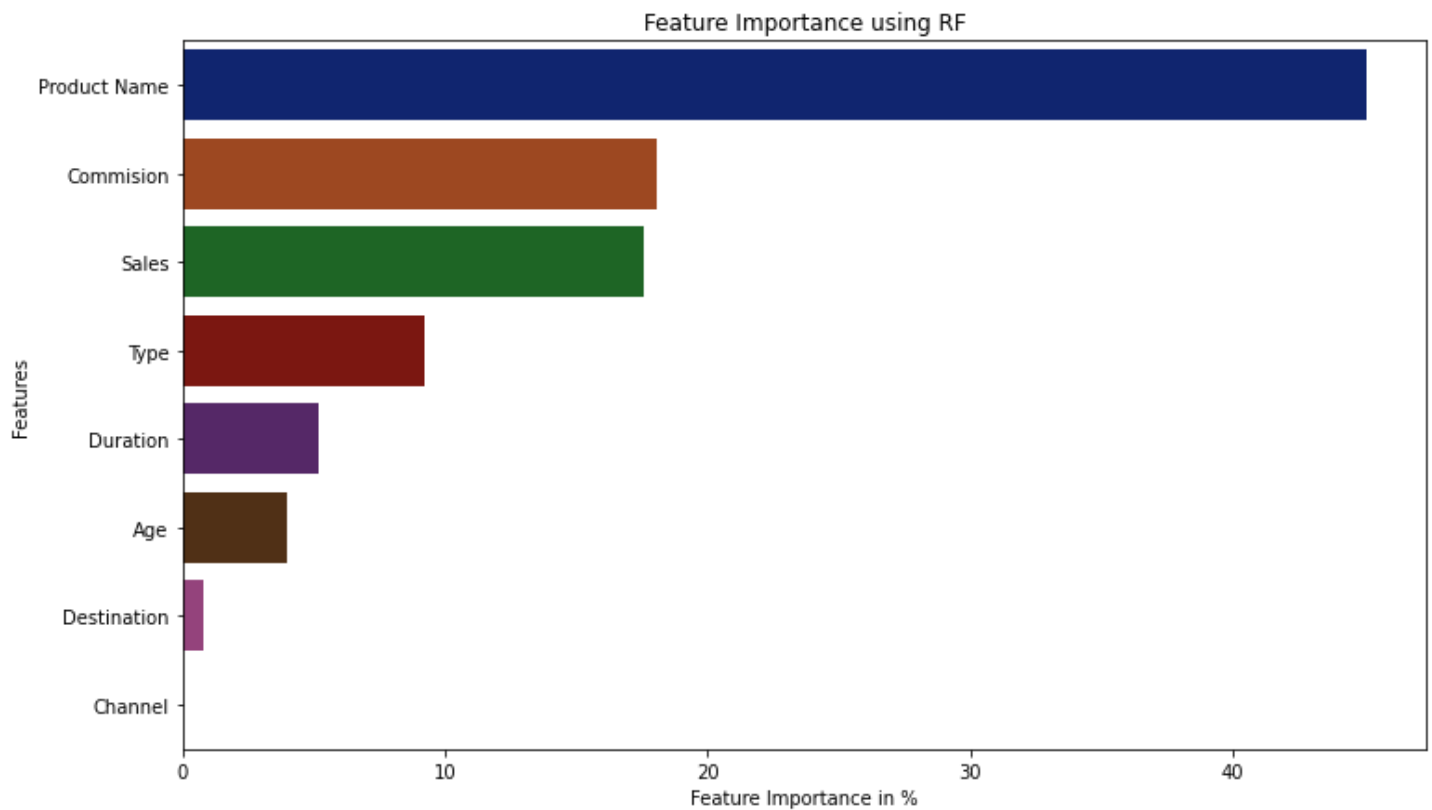


Figure 2. 7 – RANDOM FOREST FEATURE IMPORTANCE

This graph represents the features importance to build the decision tree model. Product name plays the most important roll more than 40 % to build this model followed by commission with around 19 % then sales around 18 % then rest is contributed by rest of features. Channel is the least important feature to build this model.

Creating Artificial Neural Network Model (ANN Model)

Artificial neural networks (ANNs) are biologically inspired computer programs designed to simulate the way in which the human brain processes information. ANNs gather their knowledge by detecting the patterns and relationships in data and learn (or are trained) through experience, not from programming. An ANN is formed from hundreds of single units, artificial neurons or processing elements (PE), connected with coefficients (weights), which constitute the neural structure and are organized in layers.

Building the model

Parameters used to build the model

1. Hidden layer size = 600
2. Maximum iteration = 5000
3. Solver = Stochastic Gradient Descent
4. Tolerance = 0.001

We have chosen these parameters using Grid search CV method which provides best possible combination of hyper parameters outcome using given parameters.

REPORT FOR TRAIN DATA

	precision	recall	f1-score	support
0	0.83	0.78	0.81	1471
1	0.55	0.63	0.59	629
accuracy			0.74	2100
macro avg	0.69	0.71	0.70	2100
weighted avg	0.75	0.74	0.74	2100

REPORT FOR TEST DATA

	precision	recall	f1-score	support
0	0.78	0.82	0.80	605
1	0.59	0.54	0.56	295
accuracy			0.73	900
macro avg	0.69	0.68	0.68	900
weighted avg	0.72	0.73	0.72	900

PARAMETER DESCRIPTION

HIDDEN LAYERS SIZE

In neural networks, a hidden layer is located between the input and output of the algorithm, in which the function applies weights to the inputs and directs them through an activation function as the output.

MAXIMUM ITERATIONS

It is equivalent to maximum number of epochs you want the model get trained on. It is called as maximum because the learning could get stopped before reaching the maximum number of iterations as well based on other termination criteria.

SOLVER

Solvers are one of the hyperparameters of the ANN. Solver is the algorithm used in the process of backpropagation to calculate the weights of the neural network.

TOLERANCE

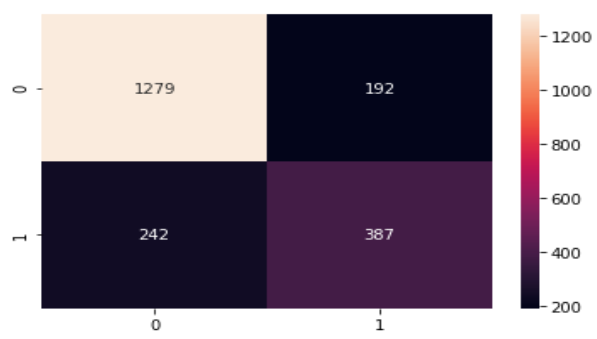
Tolerance is a hyper parameter function to keep the error in limits i.e. if the difference between two iterations goes below the tolerance level than it means further iterations is not making much difference in model performance and we can terminate the model.

Q 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:

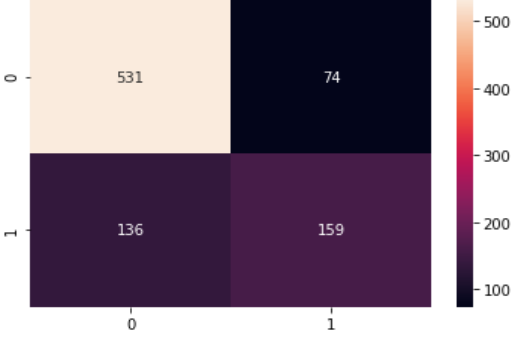
Performance report for the Decision Tree Model.

PERFORMANCE REPORT FOR TRAIN



Accuracy score- 0.7933333333333333

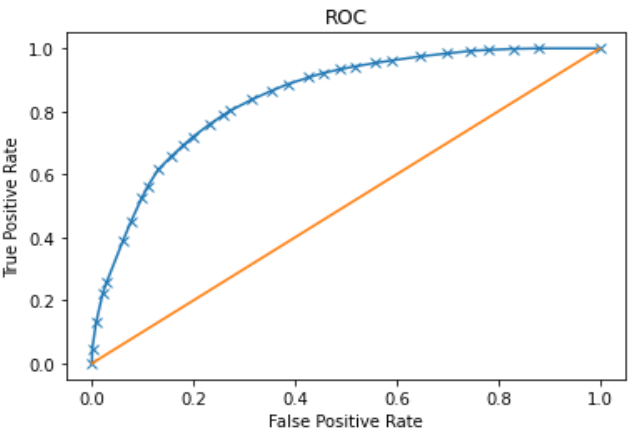
PERFORMANCE REPORT FOR TEST



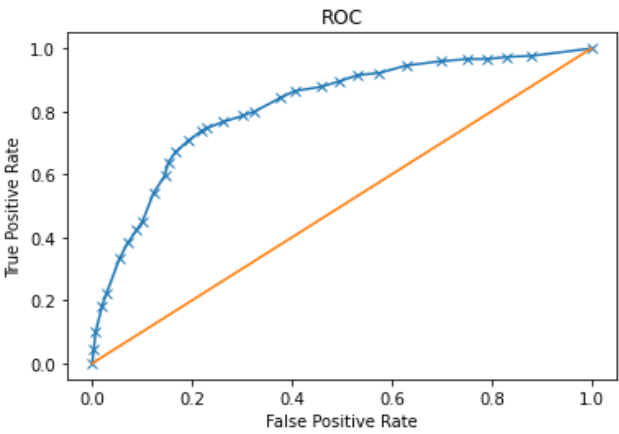
Accuracy score- 0.7666666666666667

REPORT FOR TRAIN DATA				
	precision	recall	f1-score	support
0	0.84	0.87	0.85	1471
1	0.67	0.62	0.64	629
accuracy			0.79	2100
macro avg	0.75	0.74	0.75	2100
weighted avg	0.79	0.79	0.79	2100

REPORT FOR TEST DATA				
	precision	recall	f1-score	support
0	0.80	0.88	0.83	605
1	0.68	0.54	0.60	295
accuracy			0.77	900
macro avg	0.74	0.71	0.72	900
weighted avg	0.76	0.77	0.76	900



ROC score- 0.8444813830505836



ROC score- 0.8149488723910912

For training data confusion matrix has following values

For Training Data

True Negative – 1279

True Positive – 387

False Negative - 242

False Positive - 192

For Test Data

True Negative – 531

True Positive – 159

False Negative - 136

False Positive - 74

Confusion Matrix

As we can check from the above report of the model the prediction of True Negative values in training data is higher as compare to True positive. In test data True positive and false positive values are almost similar so the prediction of the model for test data is not that good as compare to training data.

Classification Report

Same things can be seen in classification report that prediction for training dataset is Precision, Recall and F1 Score values for true negative is high 84%, 87% and 85% which is very good but for True positive prediction for training data set is not that good, 67 %,62% and 64% respectively.

For the test data set these values has dropped a bit for prediction of True Negative, Precision and Recall is 80%, 88% and 83%, for True positive its 68%,54% and 60% Respectively.

Accuracy Score

Accuracy score for both the training data set and test data set around 79% and 77%. As these values are close to each other we can say model is a robust model.

ROC Curve and Score.

The area under the curve for both training and test data set is also close and high which is 84% and 81% respectively which very good for model performance.

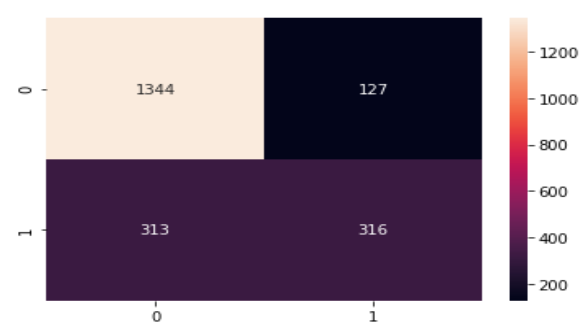
CONCLUSION

The model performance is good for prediction of True Negative (0) but not that good for True Positive (1) but the accuracy and roc score is good for the model.

Since all the values in all the reports are almost similar for both test and train data, we can say that the model is neither over-fitted nor under-fitted in any manner it is a robust model.

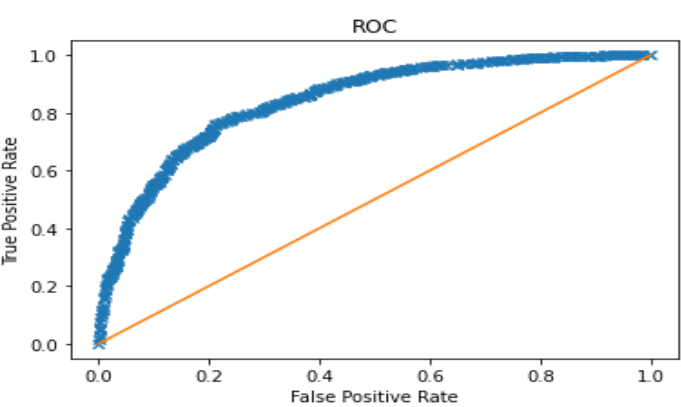
Performance Report for Random Forest Model

PERFORMANCE REPORT FOR TRAIN



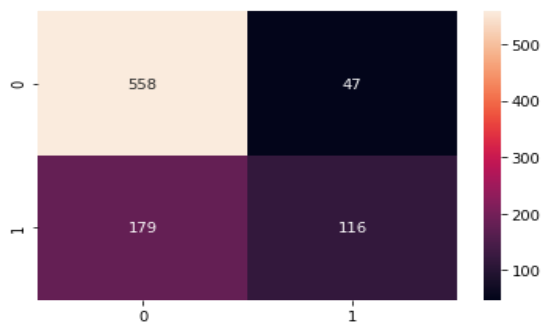
Accuracy score- 0.790952380952381

REPORT FOR TRAIN DATA					
	precision	recall	f1-score	support	
0	0.81	0.91	0.86	1471	
1	0.71	0.50	0.59	629	
accuracy			0.79	2100	
macro avg	0.76	0.71	0.72	2100	
weighted avg	0.78	0.79	0.78	2100	



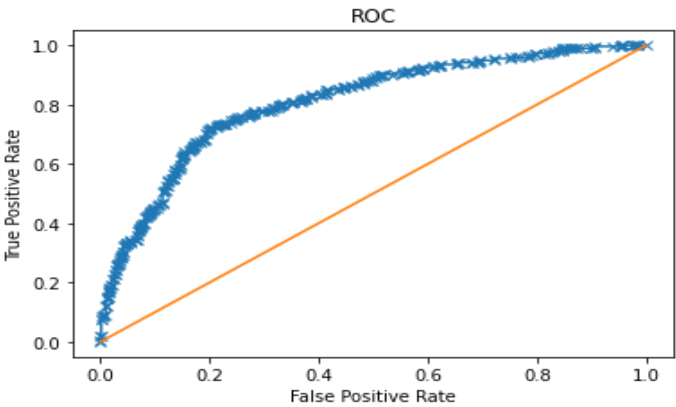
ROC score- 8428445440682013

PERFORMANCE REPORT FOR TEST



Accuracy score- 0.7488888888888889

REPORT FOR TEST DATA					
	precision	recall	f1-score	support	
0	0.76	0.92	0.83	605	
1	0.71	0.39	0.51	295	
accuracy			0.75	900	
macro avg	0.73	0.66	0.67	900	
weighted avg	0.74	0.75	0.73	900	



ROC score- 0.8136405659055891

For training data confusion matrix has following values

For Training Data

True Negative – 1344

True Positive – 316

False Negative - 313

False Positive - 127

For Test Data

True Negative – 558

True Positive – 116

False Negative - 179

False Positive - 47

Confusion Matrix

As we can check from the above report of the model the prediction of True Negative values in training data is higher as compare to True positive. In test data True positive and false positive values are almost similar so the prediction of the model for test data is not that good as compare to training data.

Classification Report

Same things can be seen in classification report that prediction for training dataset is Precision, Recall and F1 Score values for true negative is high 81%, 92% and 86% which is very good but for True positive prediction for training data set is not that good, 72 %, 52% and 59 % respectively.

For the test data set these values has dropped a bit for prediction of True Negative, Precision, Recall and F1 Score is 76% , 93% and 83%, for True positive its 72% ,39% and 50% Respectively.

As there is a significant drop in recall value in train test (more than 10%) , we can say that model is overfitted.

Accuracy Score

Accuracy score for both the training data set and test data set around 79% and 75%. As these values are close to each other we can say model is a robust model.

ROC Curve and Score.

The area under the curve for both training and test data set is also close and high which is 84% and 81% respectively which very good for model performance.

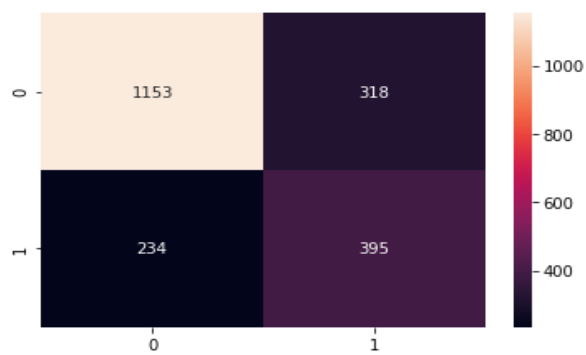
CONCLUSION

The model performance is good for prediction of True Negative (0) but not that good for True Positive (1) but the accuracy and roc score is good for the model.

Since all the values in all the reports are almost similar for both test and train data, we can say that the model is over-fitted .

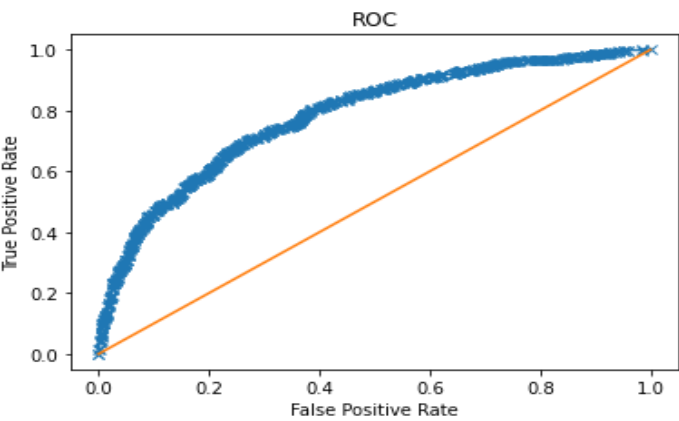
Performance Report for Artificial Neural Network Model

PERFORMANCE REPORT FOR TRAIN



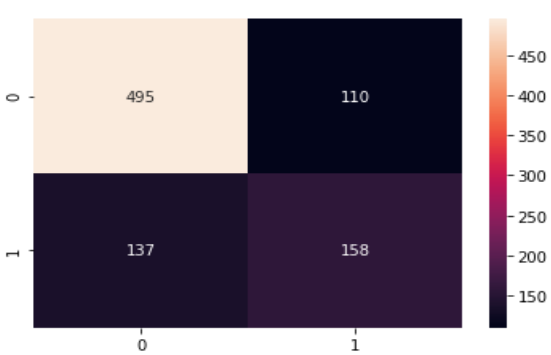
Accuracy score- 0.7371428571428571

REPORT FOR TRAIN DATA					
	precision	recall	f1-score	support	
0	0.83	0.78	0.81	1471	
1	0.55	0.63	0.59	629	
accuracy			0.74	2100	
macro avg	0.69	0.71	0.70	2100	
weighted avg	0.75	0.74	0.74	2100	



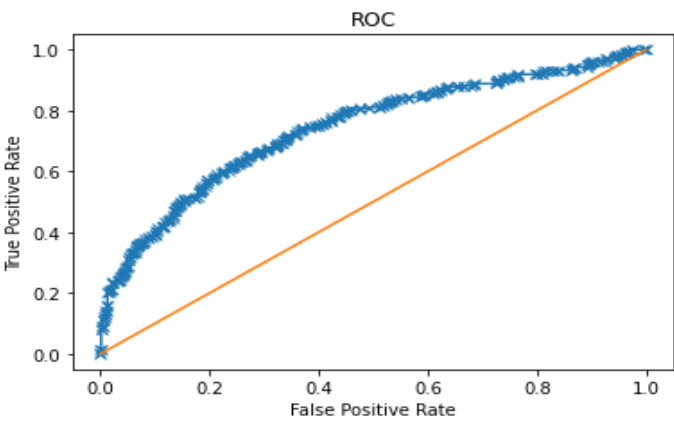
ROC score- 0.7834795446464179

PERFORMANCE REPORT FOR TEST



Accuracy score- 0.7255555555555555

REPORT FOR TEST DATA					
	precision	recall	f1-score	support	
0	0.78	0.82	0.80	605	
1	0.59	0.54	0.56	295	
accuracy			0.73	900	
macro avg	0.69	0.68	0.68	900	
weighted avg	0.72	0.73	0.72	900	



ROC score- 0.7424653312788906

For training data confusion matrix has following values

For Training Data
True Negative – 1153
True Positive – 395
False Negative - 234
False Positive - 318

For Test Data
True Negative – 495
True Positive – 114
False Negative - 137
False Positive - 110

Confusion Matrix

As we can check from the above report of the model the prediction of True Negative values in training data is higher as compare to True positive. In test data True positive and false positive values are almost similar so the prediction of the model for test data is not that good as compare to training data.

Classification Report

Same things can be seen in classification report that prediction for training dataset is Precision, Recall and F1 Score values for true negative is high 83%, 78% and 81% which is very good but for True positive prediction for training data set is not that good, 55 % and 63 % and 59% respectively.

For the test data set these values has dropped a bit for prediction of True Negative, Precision, Recall and F1 Score is 72%, 82% and 80%, for True positive its 59%, 54% and 56% Respectively.

Accuracy Score

Accuracy score for both the training data set and test data set around 74% and 73%. As these values are close to each other we can say model is a robust model.

ROC Curve and Score.

The area under the curve for both training and test data set is also close and high which is 78% and 74% respectively which very good for model performance.

CONCLUSION

The model performance is good for prediction of True Negative (0) but not that good for True Positive (1) but the accuracy and roc score is good for the model.

Since all the values in all the reports are almost similar for both test and train data we can say that the it's a robust model and is neither over-fitted nor under-fitted in any manner.

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

Solution:

MODELS PERFORMANCE COMPARISON											
		PRECISION		RECALL		F1 SCORE		ACCURACY		ROC SCORE	
		TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Decision Tree model	0	0.84	0.8	0.87	0.88	0.85	0.83	0.79	0.77	0.84	0.81
	1	0.67	0.68	0.62	0.54	0.64	0.6				
Random Forest Model	0	0.81	0.76	0.91	0.92	0.86	0.83	0.79	0.75	0.84	0.81
	1	0.71	0.71	0.5	0.39	0.59	0.51				
Artificial Neural Network	0	0.83	0.78	0.78	0.82	0.81	0.8	0.74	0.73	0.78	0.74
	1	0.55	0.59	0.63	0.54	0.59	0.56				

Table 2. 4 – MODEL COMPARISON

From the above table, we can compare the performance of each model,

PRECISION

For precision values of train data Decision tree model and ANN model has better predictions for 0's(True Negative) but for 1's random forest prediction is much better, In test data all the precision values dropped a bit from train data but decision tree model has the higher value for 0's, for 1 random forest is still has the highest value.

RECALL

For Recall random forest has leading with the highest percentage for predicting 0's i.e. 91 and 92 % but for 1's the prediction is very poor as compare to other models. There is 11% drop in train and test for random forest whereas there is only 9% and 6% drop in recall value of train and test for ANN and decision tree model.

F1 SCORE

For F1 score decision tree model performance is much better compare to other models 85% and 64% for train and 83 % and 60% of test data for 0's and 1's respectively.

ACCURACY AND ROC SCORE

For Accuracy and roc score values decision tree model performance is better than other models like random forest and ANN.

CONCLUSION.

We checked the comparison of each and every model in detail which leads to the conclusion that Decision Tree Model is best suited for this business problem and its overall performance exceeds the other models and its accuracy is highest among all three models i.e. 79 and 77 % for train and test resp.

As the company's target is to predict that which customer will raise the insurance claim(1's), we need to focus on predicting the 1's correctly and decision tree model has the best performance for that amongst all three.

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

Solution:

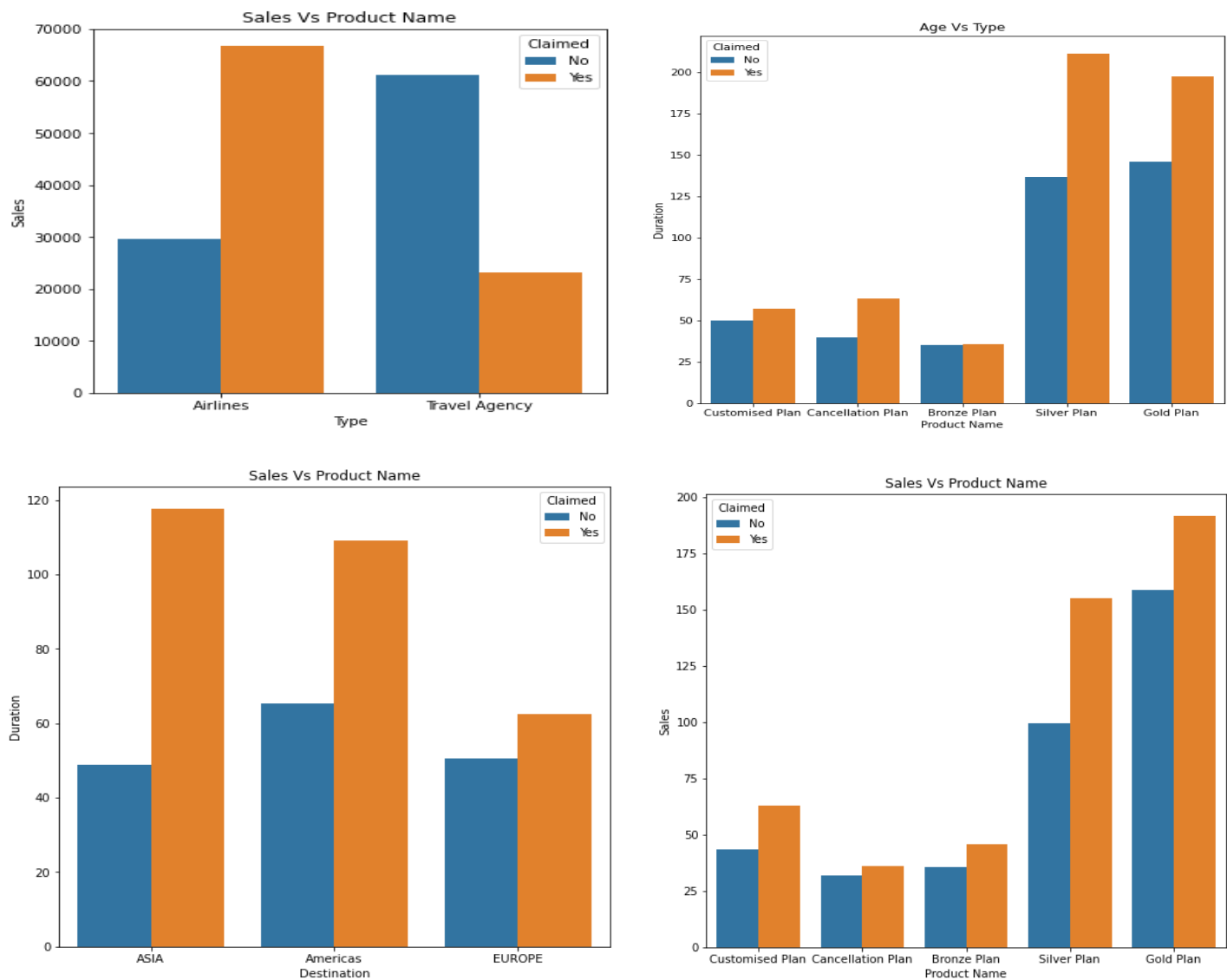


Figure 2. 8 – FEATURES RELATION FOR BUSINESS

Business Insights

Agency:

Sales for airlines and travel agency has a significant difference. Customer prefer travel agency for insurance purchase rather than airlines but still airlines has more sales in the claims raised cases as compare to travel agency.

Recommendation:

We need to understand the root cause of the problem that why insurance through airlines has too many claims and try to reduce the insurance claims by addressing the problem.

Product Name:

Customer traveling for longer duration prefers gold and silver plans for their insurance plan but these also have the highest claims raised among all other plans. As customized plans are more preferred among customers but gold and silver plans have higher sales with significant difference.

Recommendation:

We can focus on modifying these plans according to customer needs and try to reduce the claims raised by addressing the root cause.

Destination:

We can see from the plot that people who travel for the longer duration tends to raise the insurance claims more often.

Recommendation:

We Need to find the root cause and take precautions for people who are traveling for longer duration.

