DATA MINING PROJECT REPORT 2023

Sangeeth A

PGP-DSBA Online

July - 2022

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PROBLEM 1 – SUMMARY

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

INTRODUCTION

The purpose of this exercise is to conduct cluster analysis on the given dataset and develop a cluster segmentation to give promotional offers to the customers of a leading bank. We are directed to conduct hierarchical clustering as well as K-Means clustering in the given dataset and identify different clusters.

DATA DESCRIPTION

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

SAMPLE OF THE DATASET

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

Table 1 – Sample dataset

The dataset contains seven features, based on which we have to we have to do cluster segmentation of the data.

EXPLORATORY DATA ANALYSIS

The dataset contains 210 entries and 7 columns, there are no null value present in the dataset.

NO.	Column	Non – Null content	Data Type
1	spending	210	Float64
2	advance_payments	210	Float64
3	probability_of_full_payment	210	Float64
4	current_balance	210	Float64
5	credit_limit	210	Float64
6	min_payment_amt	210	Float64
7	max_spent_in_single_shopping	210	Float64

Table 2 – Exploratory Data Analysis

DESCRIPTIVE DATA 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

Table 3 – Descriptive Data Analysis

1. Probability of full payment is high.
2. Total 210 entries are there.
3. Maximum value of spending is 21.1(in 1000RS) and minimum value is 10.59(in 1 000RS).
4. Probability of full payment has a low standard deviation compared to other vari ables.

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.

1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis)(Read the data and do exploratory data analysis (3 pts). Describe the data briefly. Interpret the inferences for each (3 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.).

The problem statement is to read the data and do the initial necessary steps and conduct exploratory data analysis on the dataset provided.

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shoppin
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.55
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.14
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.14
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.18
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.83

Table 1.1.1 – Sample dataset

The above shown is the head of the dataset, the dataset contains seven columns and they are,

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

All the variables in the dataset is numerical and there are no categorical variable present in the data. Based on these features of the data we have to change the row to different clusters in order to provide promotional offers to the customers of the bank.

Now let us check the data types of the different features present in the dataset.

NO.	Column	Non – Null content	Data Type
1	spending	210	Float64
2	advance_payments	210	Float64
3	probability_of_full_payment	210	Float64
4	current_balance	210	Float64
5	credit_limit	210	Float64
6	min_payment_amt	210	Float64
7	max_spent_in_single_shopping	210	Float64

Table 1.1.2 – Data info

There are total 210 entries and 7 columns in the dataset, all the dataset is in float64 data type.

Now we can check for the null values present in the dataset,

NO.	Column	Null value
1	spending	0
2	advance_payments	0
3	probability_of_full_payment	0
4	current_balance	0
5	credit_limit	0
6	min_payment_amt	0
7	max_spent_in_single_shopping	0

Table 1.1.3 – Null value check

There is no null value present in the data set.

	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

Table 1.1.4 – Summary statistics

The above shown table shows the summary statistics of the dataset, the probability of full payment has low standard deviation compared to other features of the dataset. The maximum sending in a single shopping is 6.5 thousand and the minimum spending in a single shopping is 4.5 thousand.

Now let us check the skewness of the dataset,

NO.	Column	Skewness
1	spending	0.399889
2	advance_payments	0.386573
3	probability_of_full_payment	-0.537954
4	current_balance	0.525482
5	credit_limit	0.134378
6	min_payment_amt	0.401667
7	max_spent_in_single_shopping	0.561897

Table 1.1.5 – Skewness

If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed. If the skewness is less than -1 or greater than 1, the data are highly skewed. Here the skewness value of data is range from -0.5 to 0.5 therefore the data is fairly symmetrical.

Next let us check the outlier proportion of the given data,

NO.	Column	Number of outlier	Outlier proportion
1	spending	0	0
2	advance_payments	0	0
3	probability_of_full_payment	3	0.014
4	current_balance	0	0
5	credit_limit	0	0
6	min_payment_amt	2	0.009
7	max_spent_in_single_shopping	0	0

Table 1.1.6 –Outlier proportion

The above table shows the number of outlier present in the dataset and the outlier proportion ration. There are total 5 outlier present in the dataset, 3 values in probability of full payment and 2 values in minimum payment amount. When we check the outlier proportion ratio of these values probability of full payment has 0.014 and minimum payment amount has 0.009. So we can conclude that the outlier proportion ratio is very low.

EXPLORATORY DATA ANALYSIS

Now let us do the exploratory data analysis of the given data.

UNIVARIATE ANALYSIS

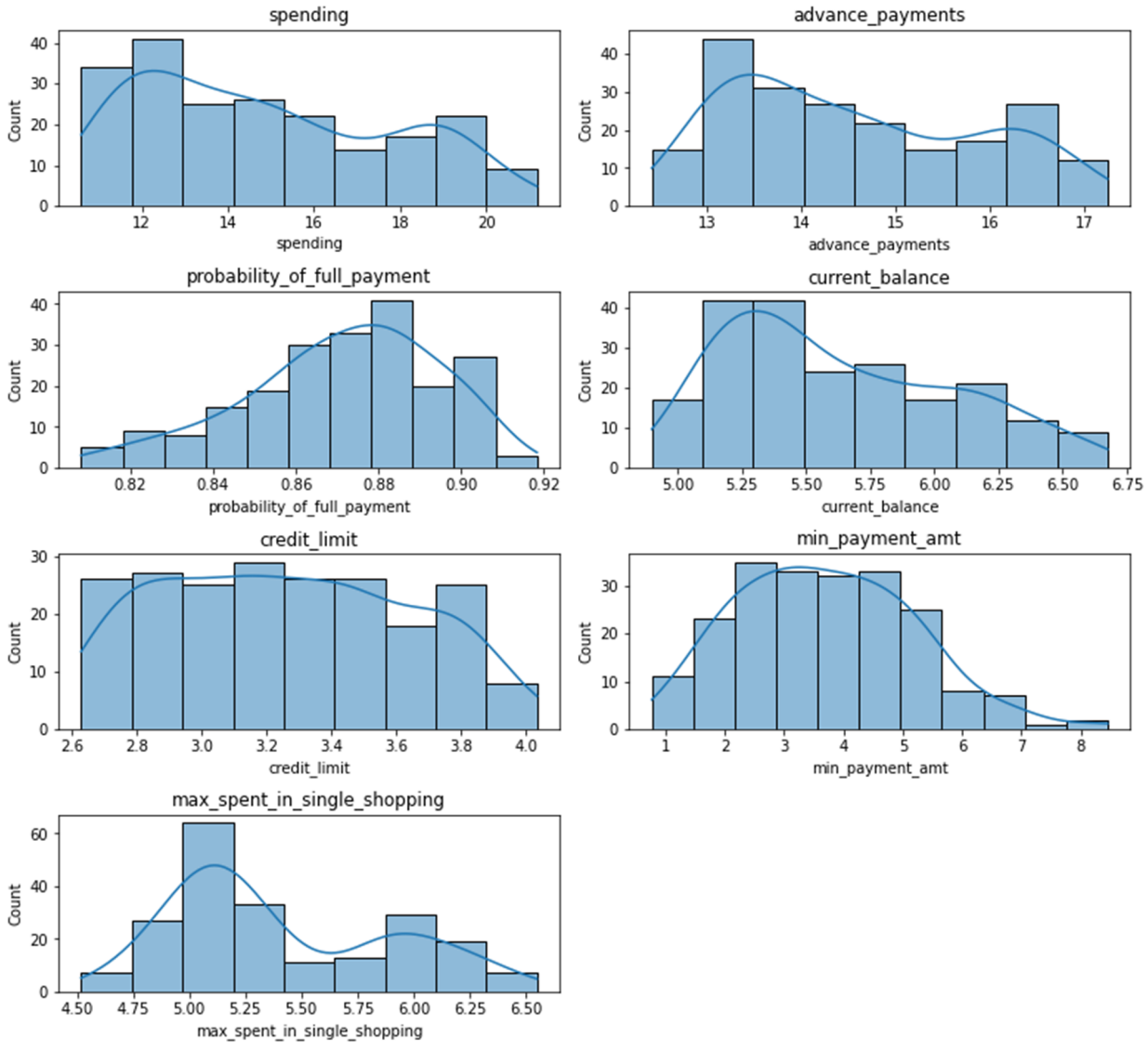

Fig 1.1.1 –Univariate analysis (HIst Plot)

The above shown is the hist plot of the different features of the dataset. The observations from the hist plot is given below.

1. Most of the people spending a maximum amount of twelve thousand rupees and very few people are spending the maximum spending amount of twenty thousand compared to the other amounts.
2. Most of the people are paying an advance payment of thirteen hundred rupees.

3. In the case of probability of full payment out of the total number around 88% of the people are expected to pay full amount.
4. Most of the people having a current balance of around 5.25 to 5.5 thousand rupees. People having a very high current balance is very less.
5. Most number of people having a credit limit of 32,000, very few people have credit limit in the range of forty thousand.
6. Most of the people pay a minimum payment amount of 200 to 300Rs, very few people pay a minimum payment amount of around 750Rs.
7. Most of the people spent a maximum amount of 5000Rs in a single shopping.

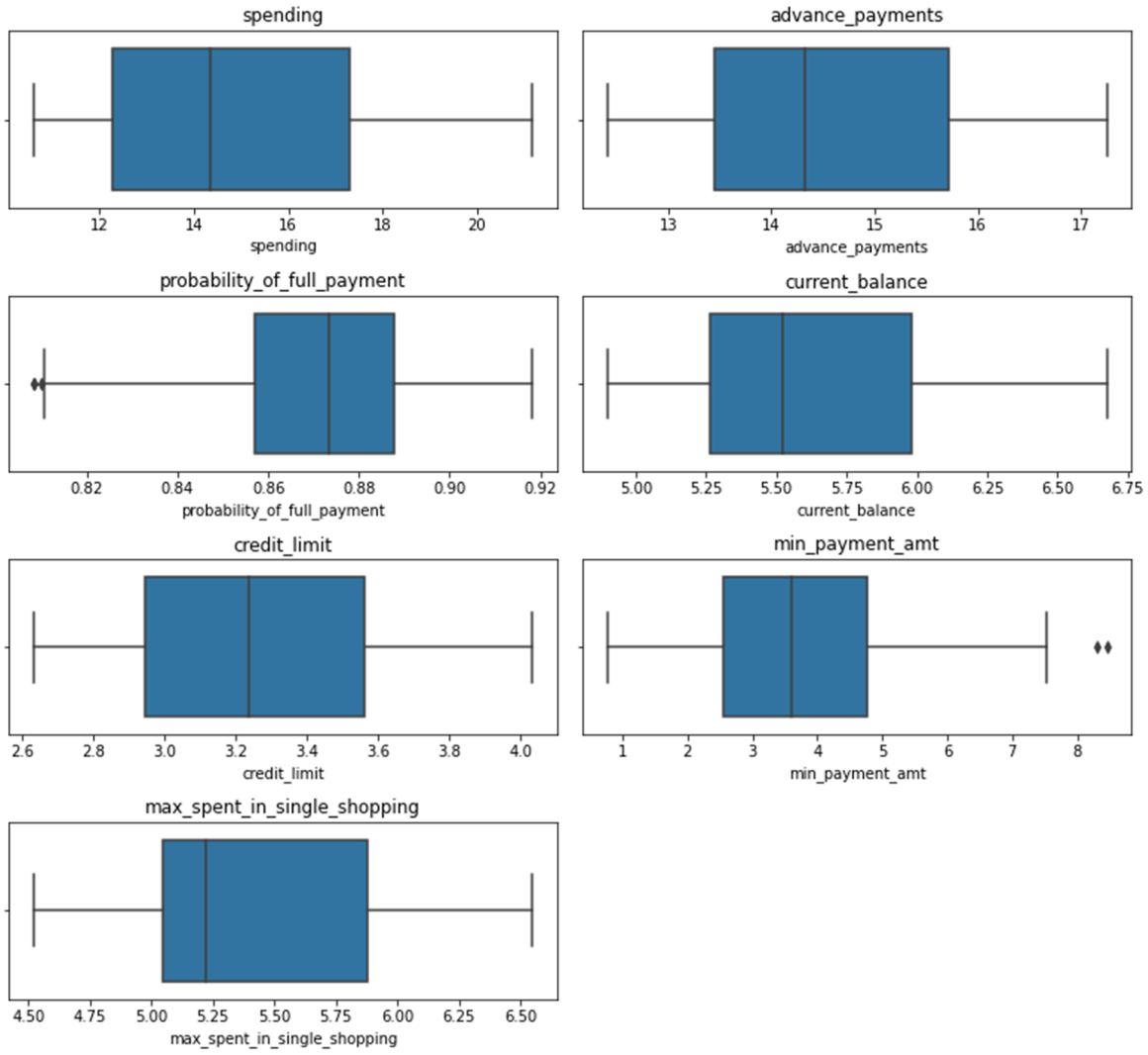

Fig 1.1.2 –Univariate analysis (Box Plot)

The above shown is the box plot of different features in the dataset. The observations from the dataset is below,

1. Spending data is right skewed, the median value is in between fourteen thousand and fifteen thousand.
2. Advance payment is also right skewed, and the median value is in between fourteen hundred and fifteen hundred.
3. Probability of full payment data is left skewed, and the median of the data is lies in between 0.87 and 0.88. There are outliers present in the probability of full payment data.
4. The current balance data is right skewed data, and the median value is about 5.5 thousand.
5. Credit limit data is right skewed, and the median value lies in between 32 thousand and 33 thousand.
6. Minimum payment amount data is right skewed, and the median value lies in between 300 and 400. There are outlier present in the data.
7. Maximum spend in single shopping data is right skewed, and the median value lies in between 5 and 5.25 thousand.

BIVARIATE ANALYSIS

Now let us do the bivariate analysis of the data, since all the entries are numerical in the dataset, we can go for scatter plot or lm plot analysis of the data.

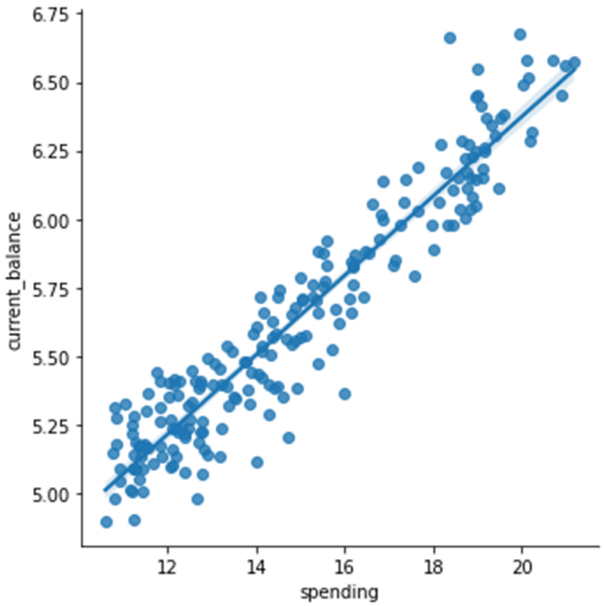

Fig 1.1.3 –Bivariate analysis (Spending x Current balance)

The figure represent the relation between the spending and the current balance of the customer, from the figure it is very clear that the spending is directly related to

the current balance of the customer. As the current balance increases the spending also increase and vice versa.

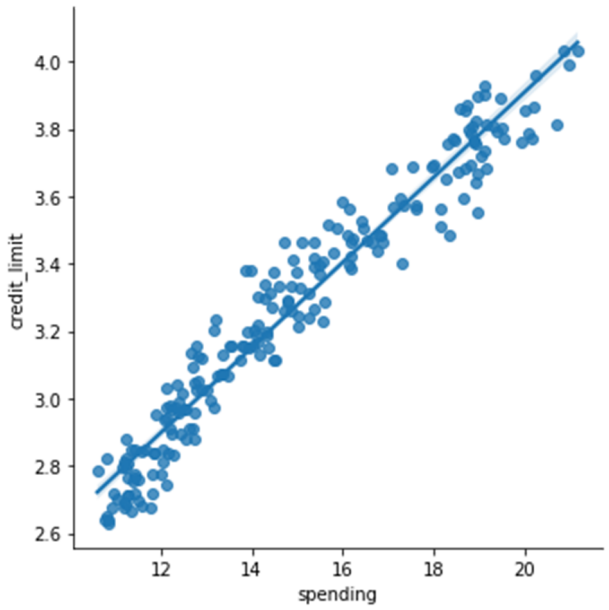

Fig 1.1.4 –Bivariate analysis (Spending x Credit limit)

The figure represents the relation between the spending and the credit limit of the customer. The credit limit and spending is directly related, as the credit limit increases the spending also increases.

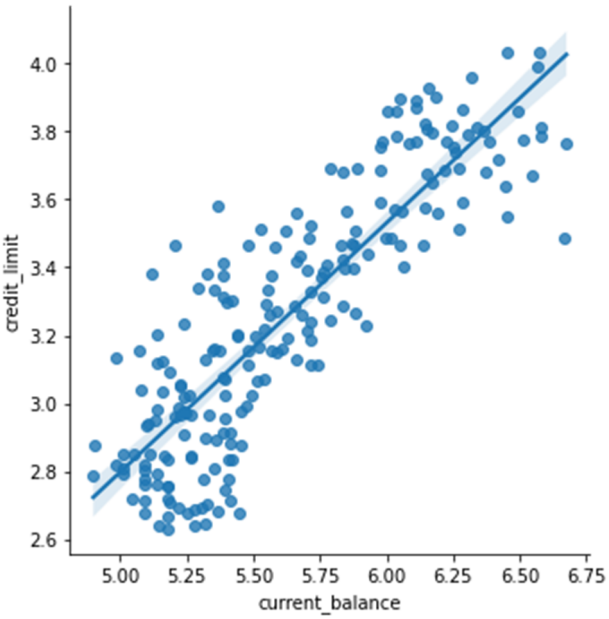

Fig 1.1.5 –Bivariate analysis (Current balance x Credit limit)

Above shown figure represents the relation between the current balance and credit limit. The current balance and credit limit directly related as the current balance increases the credit limit also increases.

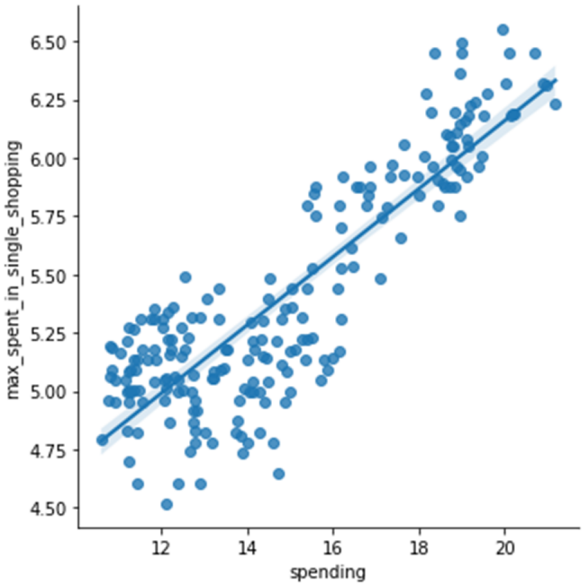

Fig 1.1.6 –Bivariate analysis (Spending x Maximum spent in single shopping)

The figure shows the relation between the spending and maximum spent in single shopping of the customers. From the figure it is clear that these data's are highly correlated.

MULTIVARIATE ANALYSIS

Fig 1.1.7 –Multivariate analysis (Heatmap)

The figure represents the correlation heatmap between different features of the dataset. From the dataset it is clear that there is high correlation between spending and advance payment, spending and current balance, spending and credit limit, advance payment and credit limit, advance payment and current balance, spending and credit limit, spending and current balance.

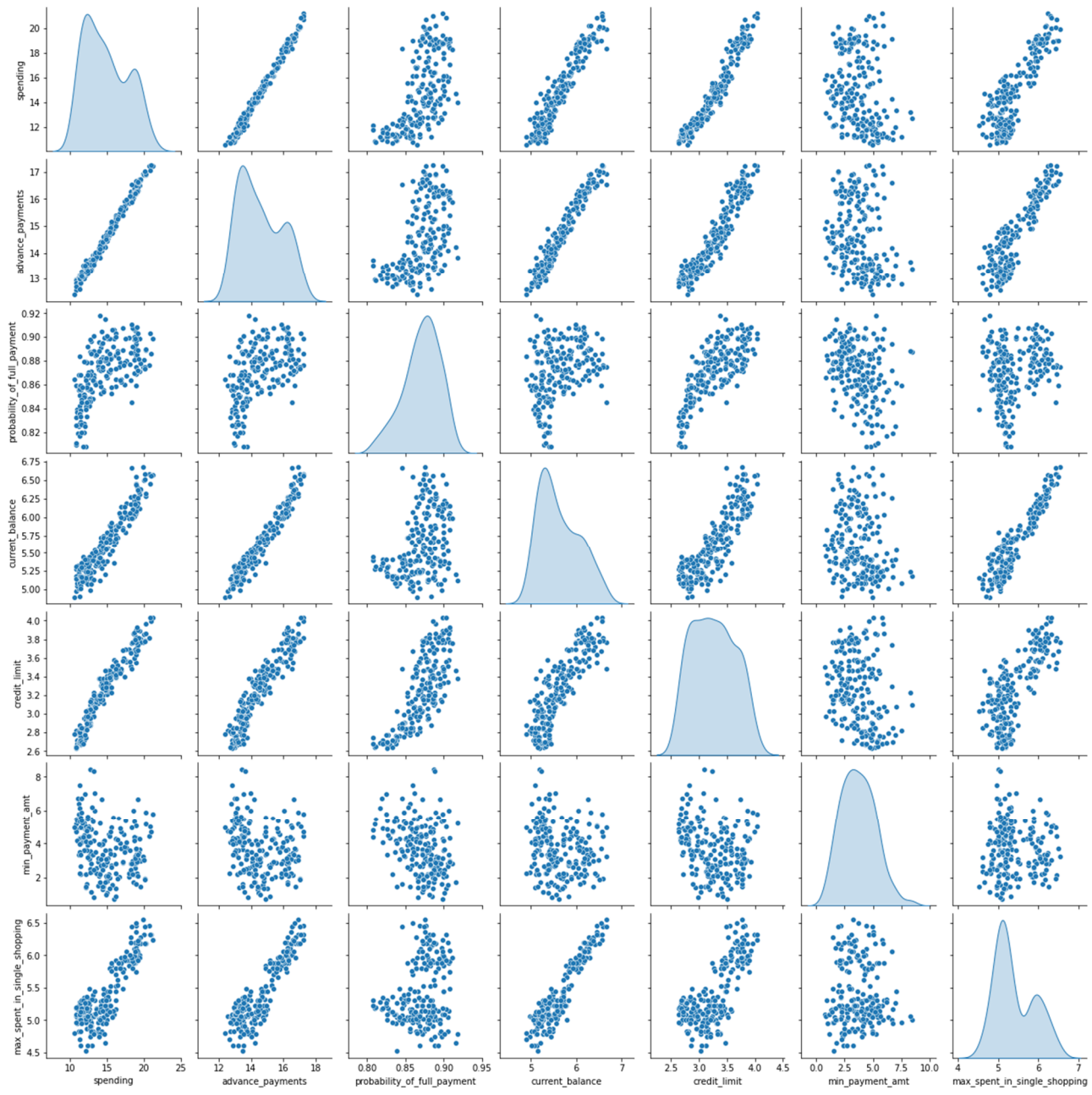

Fig 1.1.8 –Multivariate analysis (Pairplot)

Pairplot is similar to heatmap, it shows the correlation between the features of the dataset. As mentioned in the findings of the heat map pair plot also shows the high correlation between spending and advance payment, spending and current balance, spending and credit limit, advance payment and credit limit, advance payment and current balance, spending and credit limit, spending and current balance.

1.2Do you think scaling is necessary for clustering in this case? Justify. (Do you think scaling is necessary for clustering in this case? Justify The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling and which method is he/she using to do the scaling. Can also comment on how that method works.).

In this case the scaling of data is necessary, because if we check the data description the given details are provided in different multiples of ten. For example spending in thousands, advance payments in hundreds, probability of full payment in ratio, current balance in thousands, credit limit in ten thousands, minimum payment amount in hundreds and maximum spent in single shopping in thousands. If we plot the data we will understand it clearly.

Fig 1.2.1 –Data before scaling

	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

Table 1.2.1 –Data before scaling

If we check the figure and table the data is distributed to various levels and not uniform distribution. In the case of clustering it will use Euclidean distance for calculation, which is very sensitive, so this will affect the clustering process

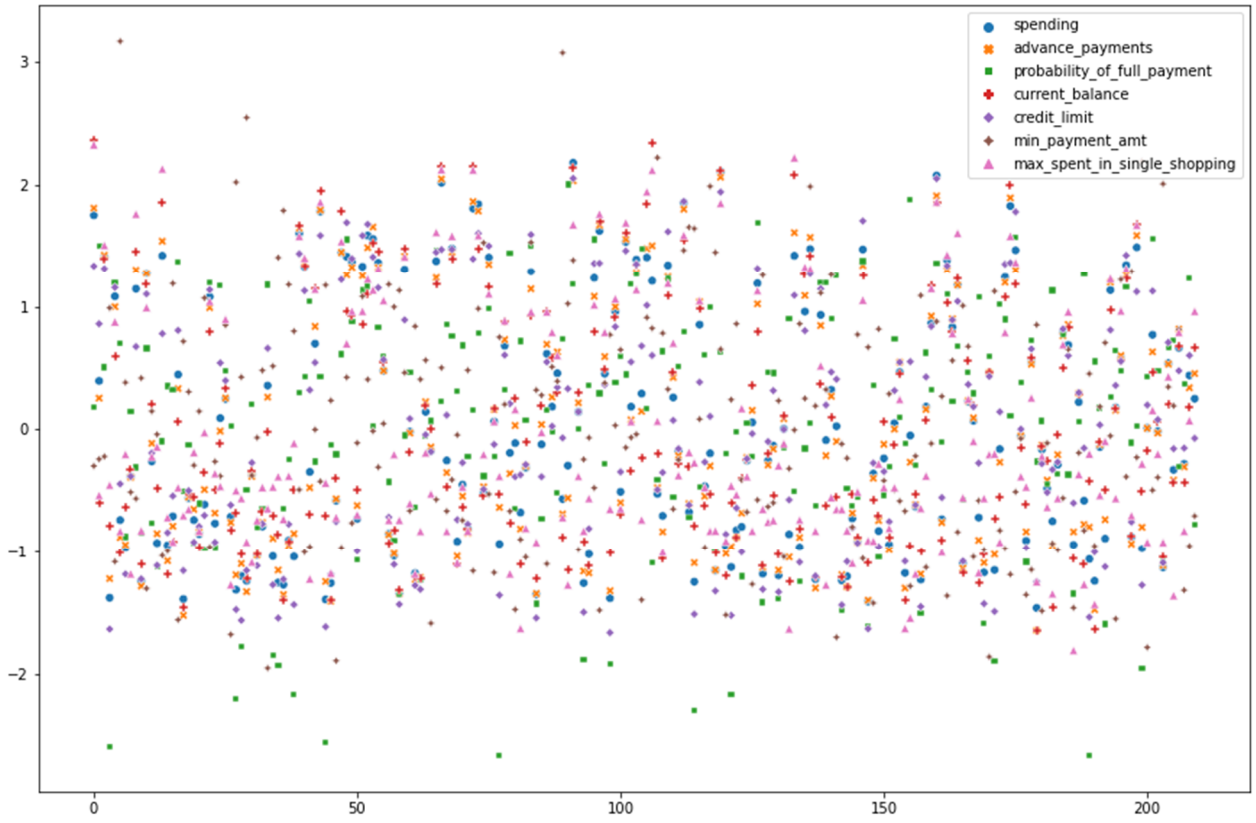

Fig 1.2.2 –Data after scaling

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.178230	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	1.501773	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	0.504874	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-2.591878	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	1.196340	0.591544	1.155464	-1.088154	0.874813

Table 1.2.2 –Data after scaling

The figure and table shows the data after the scaling process, we will use z score method or standard scaler method for the scaling of data. Now all the values are ranging from -2 to 2.

Now let us check the standard deviation and the variance of the data before and after the scaling process,

NO.	Column	Standard devation
1	spending	2.909699
2	advance_payments	1.305959
3	probability_of_full_payment	0.023629
4	current_balance	0.443063
5	credit_limit	0.377714
6	min_payment_amt	1.503557
7	max_spent_in_single_shopping	0.491480

Table 1.2.3 –Standard deviation before scaling

NO.	Column	Standard devation
1	spending	1.002389
2	advance_payments	1.002389
3	probability_of_full_payment	1.002389
4	current_balance	1.002389
5	credit_limit	1.002389
6	min_payment_amt	1.002389
7	max_spent_in_single_shopping	1.002389

Table 1.2.4 –Standard deviation after scaling

NO.	Column	Variance
1	spending	8.466351
2	advance_payments	1.705528
3	probability_of_full_payment	0.000558
4	current_balance	0.196305
5	credit_limit	0.142668
6	min_payment_amt	2.260684
7	max_spent_in_single_shopping	0.241553

Table 1.2.5 –Variance before scaling

NO.	Column	Variance
1	spending	1.004785
2	advance_payments	1.004785
3	probability_of_full_payment	1.004785
4	current_balance	1.004785
5	credit_limit	1.004785
6	min_payment_amt	1.004785
7	max_spent_in_single_shopping	1.004785

Table 1.2.6 –Variance after scaling

The tables shows the values of standard deviation and variance of the data before and after the scaling process, before scaling process the standard deviation and variance is different for different features of the dataset and after scaling process all the values of standard deviation and variance is same for all the features in the dataset. This is why scaling is necessary for clustering process.

1.3 Apply hierarchical clustering to scaled data (3 pts). Identify the number of optimum clusters using Dendrogram and briefly describe them (4). Students are expected to apply hierarchical clustering. It can be obtained via Fclusters or Agglomerative Clustering. Report should talk about the used criterion, affinity and linkage. Report must contain a Dendrogram and a logical reason behind choosing the optimum number of clusters and Inferences on the dendrogram. Customer segmentation can be visualized using limited features or whole data but it should be clear, correct and logical. Use appropriate plots to visualize the clusters.

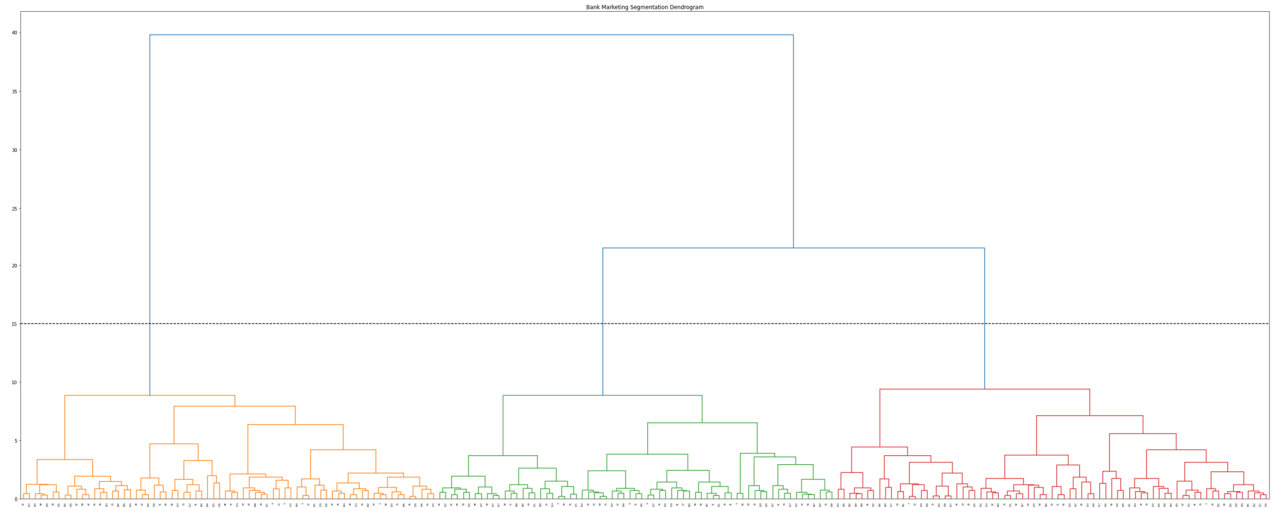

FIg 1.3.1 –Dendrogram full

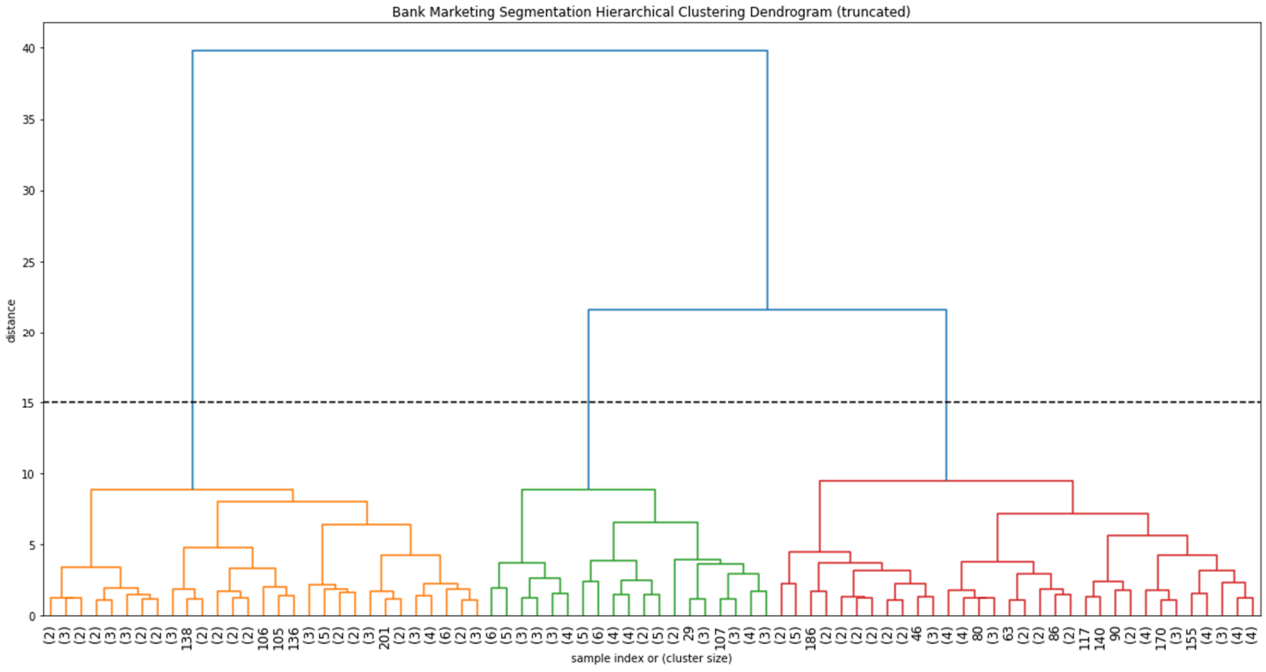

FIg 1.3.2 –Dendrogram Truncated

In the above figures, the first figure shows the complete dendrogram of the entire dataset and the second one shows the truncated type of the same one which is easy to understand. As mentioned earlier we are using scaled dataset for this type of clustering process. This type of clustering using dendrogram is known as hierarchical clustering. From the above dendrogram the vertical distance is higher for k=3. So we can conclude that optimum number of cluster for this is number 3. Next we can use fcluster method to find out which row fall into which cluster.

The linkage method we use here is ward linkage, which joins records and clusters together progressively to produce larger and larger clusters, but operates slightly differently from the general approach. The criterion we use in fcluster is maxclust, which Finds a minimum threshold r so that the cophenetic distance between any two original observations in the same flat cluster is no more than r and no more than *t* flat clusters are formed. Normally we use Euclidean distance as metric in hierarchical clustering method.

The obtained cluster distribution as per fcluster is shown below,

```
array([1, 3, 1, 2, 1, 2, 2, 3, 1, 2, 1, 3, 2, 1, 3, 2, 3, 2, 3, 2, 2, 2,
       1, 2, 3, 1, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1,
       2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 3, 3, 1,
       1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 3, 3, 1,
       1, 2, 3, 1, 3, 2, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 3, 1, 1, 2, 2, 1,
       3, 3, 1, 2, 2, 1, 3, 3, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3,
       3, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 2, 1, 2, 3, 2, 3, 2, 3, 3,
       3, 3, 3, 2, 3, 1, 1, 2, 1, 1, 1, 2, 1, 3, 3, 3, 3, 2, 3, 1, 1, 1,
       3, 3, 1, 2, 3, 3, 3, 3, 1, 1, 3, 3, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2,
       1, 2, 3, 1, 3, 2, 1, 3, 1, 3, 1, 3], dtype=int32)
```

Table 1.3.1 –Cluster allocation

Now let us import these cluster into our dataframe,

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster_1
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144	3
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837	1

Table 1.3.2–Cluster allocation into data

Let us check the pair plot of the dataframe as per the cluster segmentation and check the data distribution as per the clusters,

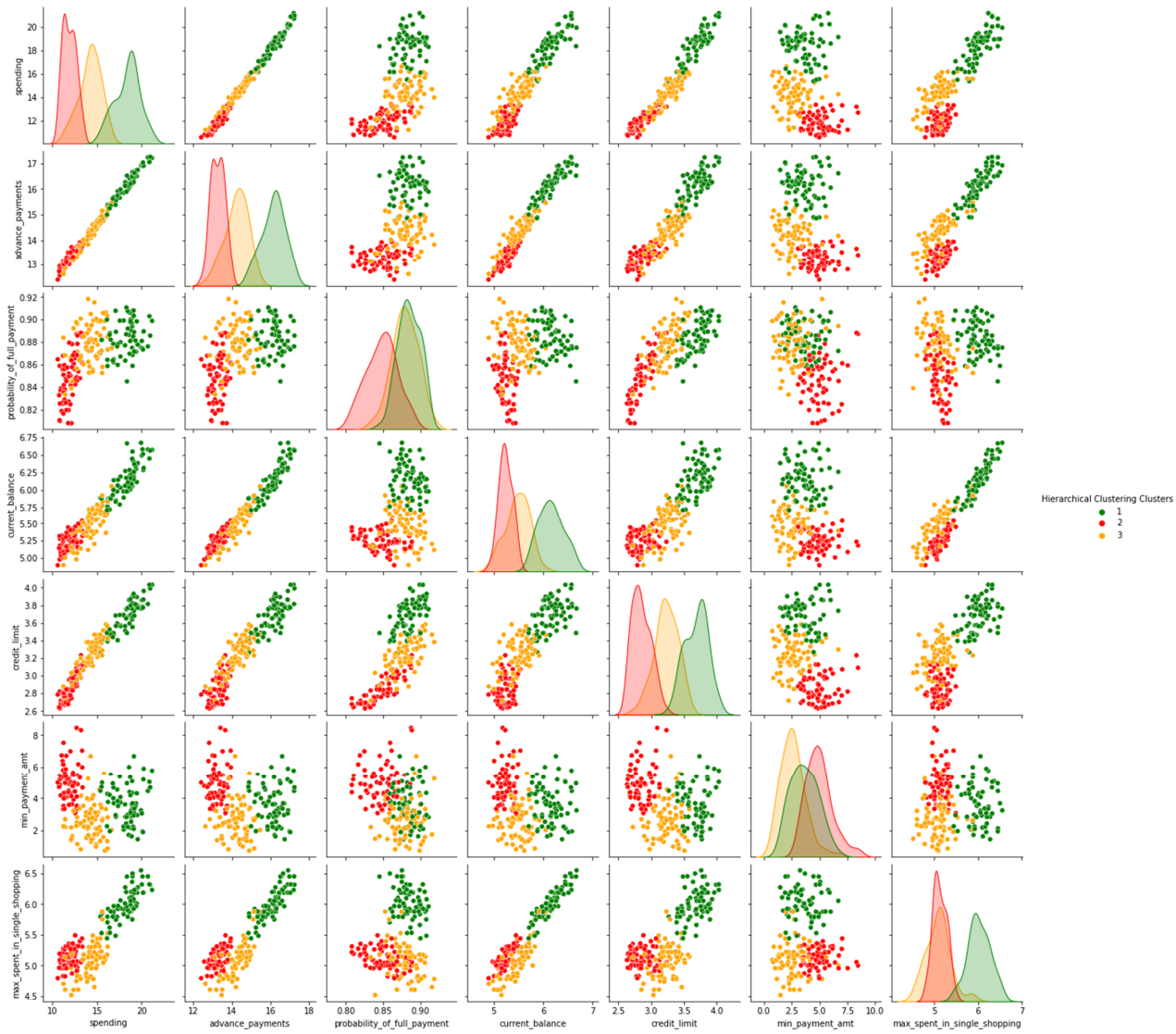

FIg 1.3.3 –Pair plot (Hierarchical Clustering)

cluster_1	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster count
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	70
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209	67
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178	73

Table 1.3.3–Groupby data as per cluster

Consider the above shown table as well as the pair plot, the table represents the mean value of different features as per the cluster segmentation and the figure represents the data distribution as per the cluster segmentation.

From the table the cluster 1 contains the group of people who spends most, total seventy values fall into cluster 1. The group cluster 1 has a mean spending of 18.37,

they have an average credit limit of 3.68 and average current balance of 6.15. The average probability of full payment of cluster 1 is 0.88 and average maximum spent of single shopping is 6.017.

In the case of cluster 2, it contains the group of people who spent least. The average spending of cluster 2 is 11.87. The average probability of full payment of cluster 2 is 0.848 which is less compared to other two clusters. The average credit limit of cluster 2 is 2.84, which is low compared to other two clusters. But in the case of minimum payment amount cluster 2 has an average 4.94 which is more compare to other two clusters. There are a total of 67 values fall into cluster 2. In the case of average maximum spending in a single shopping cluster 2 is more than cluster 3.

Now let us check the case of cluster 3, cluster 3 has a total member of 73, which is more compared to other two clusters. Cluster 3 contains the group of people who spent moderately. The average values of different features of cluster 3 is also median to other two clusters. But in the case of average minimum payment amount cluster 3 has very less average value compared to other two clusters.

Next we can check the pairplot distribution of the data as per the cluster segmentation, In the pairplot the green colour represents the group of people who spends most, yellow represents the people who spent moderately and red represents the people who spent least. From the plot itself it is clear that we should give more attention to the people in red, and should try to increase the spending of these people by providing attractive offers.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters (2 pts). Apply elbow curve (3 pts). Interpret the inferences from the model (2.5 pts). K-means clustering code application with different number of clusters. Calculation of WSS(inertia for each value of k) Elbow Method must be applied and visualized with different values of K. Reasoning behind the selection of the optimal value of K must be explained properly. Report must contain logical and correct explanations for choosing the optimum clusters using the elbow method. Append cluster labels obtained from K-means clustering into the original data frame. Customer Segmentation can be visualized using appropriate graphs.

Now let us check the K Means clustering of the same data. Let us simply apply clustering on the scaled data and check.

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

Table 1.4.1–K menas clustering When n_cluster=3

```
array([1, 0, 1, 3, 1, 3, 3, 0, 1, 3, 1, 0, 3, 1, 0, 3, 0, 3, 0, 3, 3, 3,
       1, 3, 0, 2, 0, 3, 3, 3, 0, 3, 3, 0, 3, 3, 3, 3, 3, 1, 1, 0, 2, 1,
       3, 3, 0, 1, 1, 1, 3, 1, 1, 1, 1, 2, 3, 3, 3, 1, 0, 3, 3, 2, 0, 1,
       1, 0, 1, 0, 0, 3, 1, 1, 3, 1, 0, 3, 2, 0, 0, 0, 0, 1, 3, 2, 2, 2,
       2, 3, 0, 1, 0, 3, 0, 1, 1, 2, 3, 2, 0, 1, 2, 1, 0, 1, 1, 3, 0, 1,
       2, 0, 1, 3, 3, 2, 0, 0, 3, 1, 0, 3, 3, 3, 0, 0, 1, 3, 0, 0, 3, 0,
       0, 1, 3, 1, 1, 3, 2, 0, 2, 0, 3, 3, 0, 3, 1, 3, 0, 3, 0, 3, 0, 2,
       0, 0, 0, 3, 0, 2, 1, 3, 1, 2, 1, 3, 2, 0, 0, 3, 0, 3, 0, 1, 1, 1,
       0, 0, 2, 3, 0, 0, 0, 0, 2, 2, 0, 2, 0, 3, 0, 0, 3, 1, 0, 2, 1, 3,
       1, 3, 0, 2, 0, 3, 2, 0, 2, 0, 2, 2])
```

Table 1.4.2–K menas clustering When n_cluster=4

It is very difficult to understand the optimum number of clusters by simply changing the input values, so let us calculate the inertia of the corresponding clusters and check the drop in inertia value.

```
The WSS value for 2 clusters is 1470.0000000000002
The WSS value for 3 clusters is 659.1717544870406
The WSS value for 4 clusters is 430.6589731513006
The WSS value for 5 clusters is 371.30172127754213
The WSS value for 6 clusters is 327.96082400790317
The WSS value for 7 clusters is 290.590030596822
The WSS value for 8 clusters is 264.8315308747815
The WSS value for 9 clusters is 240.68372595015978
The WSS value for 10 clusters is 220.85285825594738
```

Table 1.4.3–Inertia value corresponding to different clusters.

The above table represents the inertia value (WSS – Within sum of square value) corresponding to different number of clusters. From the table it is clear that after cluster number 3 the drop in inertia value is not significant. Let us plot the value of the inertia for more clarity.

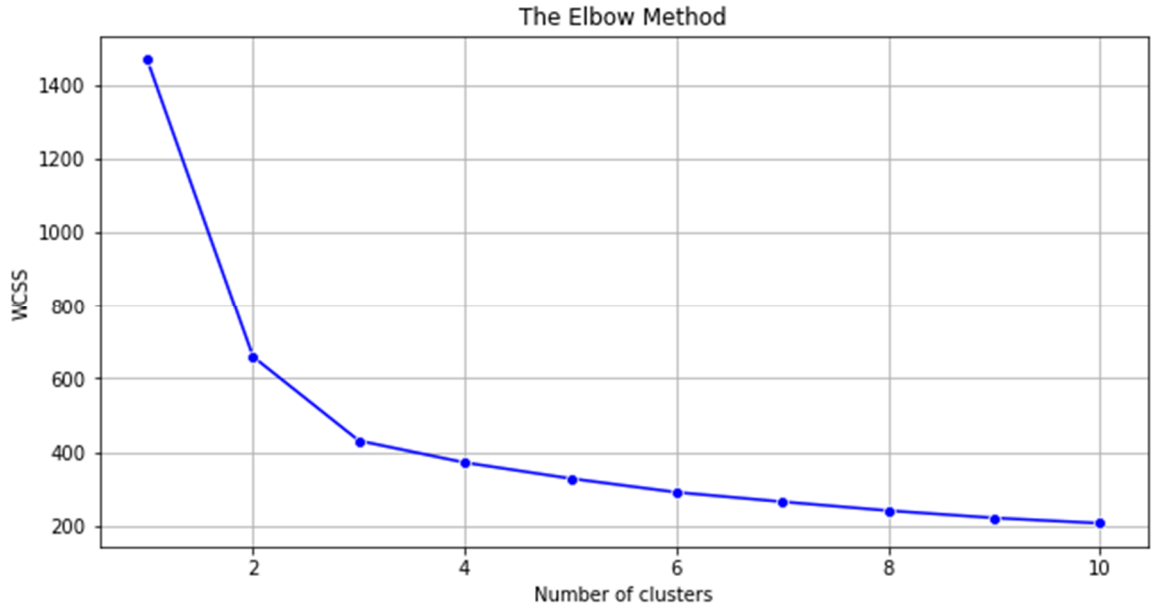

FIg 1.4.1–The elbow curve

We can notice that from k=1 to k=2 there is a significant drop, similarly from k=2 to k=3 there is also a significant drop. But after k=3 onwards the drop is gradual, the within sum of square not significantly dropping after k=3. Therefore we can conclude that k=3 is the optimum number of clusters.

So let us extract the labels corresponding to k=3 and attach it to our data.

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

Table 1.4.4–Kmeans cluster

Next we can analyse the imported K-means clusters using pair plot and we can check the average values of features with respect to clusters.

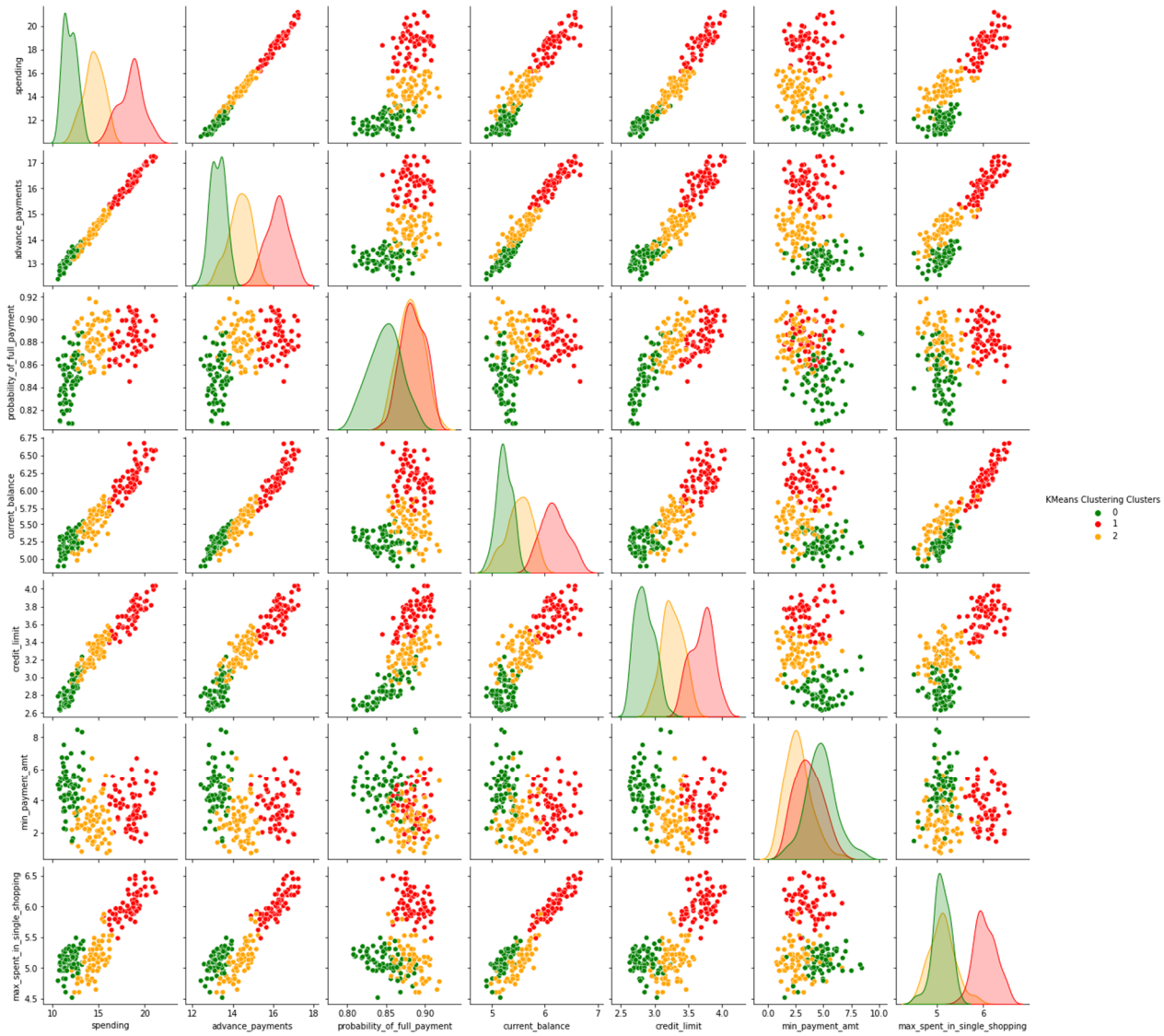

FIg 1.4.2–K-means Pairplot

KMeans Clustering Clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster count
0	11.856944	13.247778	0.848253	5.231750	2.849542	4.742389	5.101722	72
1	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	67
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	71

Table 1.4.5– Groupby data as per cluster

Consider the above shown table as well as the pair plot, the table represents the mean value of different features as per the cluster segmentation and the figure represents the data distribution as per the cluster segmentation.

From the table the cluster 0 contains the group of people who spends least, total seventy two values fall into cluster 0. The group cluster 0 has a mean spending of 11.85, they have an average credit limit of 2.84 and average current balance of 5.23. The average probability of full payment of cluster 0 is 0.84 and average maximum spent of single shopping is 5.101.

In the case of cluster 1, it contains the group of people who spent most. The average spending of cluster 1 is 11.49. The average probability of full payment of cluster 1 is 0.884 which is high compared to other two clusters. The average credit limit of cluster 1 is 3.69. But in the case of minimum payment amount cluster 2 has an average 3.63. There are a total of 67 values fall into cluster 1. In the case of average maximum spending in a single shopping cluster 1 is more than other two clusters.

Now let us check the case of cluster 2, cluster 2 has a total member of 71. Cluster 2 contains the group of people who spent moderately. The average values of different features of cluster 2 is also median to other two clusters. But in the case of average minimum payment amount cluster 2 has very less average value compared to other two clusters.

Next we can check the pairplot distribution of the data as per the cluster segmentation, In the pairplot the green colour represents the group of people who spends least, yellow represents the people who spent moderately and red represents the people who spent most. From the plot itself it is clear that we should give more attention to the people in green, and should try to increase the spending of these people by providing attractive offers.

1.5 Describe cluster profiles for the clusters defined (2.5 pts). Recommend different promotional strategies for different clusters in context to the business problem in-hand (2.5 pts). After adding the final clusters to the original dataframe, do the cluster profiling. Divide the data in the finalyzed groups and check their means. Explain each of the group briefly. There should be at least 3-4 Recommendations. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks will only be allotted if the recommendations are correct and business specific. variable means. Students to explain the profiles and suggest a mechanism to approach each cluster. Any logical explanation is acceptable.

cluster_1	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster count
1	18.371429	16.145429	0.884400	6.158171	3.684629	3.639157	6.017371	70
2	11.872388	13.257015	0.848072	5.238940	2.848537	4.949433	5.122209	67
3	14.199041	14.233562	0.879190	5.478233	3.226452	2.612181	5.086178	73

Table 1.5.1– Groupby data as per hierarchical cluster

KMeans Clustering Clusters	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	cluster count
0	11.856944	13.247778	0.848253	5.231750	2.849542	4.742389	5.101722	72
1	18.495373	16.203433	0.884210	6.175687	3.697537	3.632373	6.041701	67
2	14.437887	14.337746	0.881597	5.514577	3.259225	2.707341	5.120803	71

Table 1.5.2– Groupby data as per K - Means cluster

The above two table represents the cluster segmentation of whole data based on the average of the features. In hierarchical clustering cluster1 is the high spending group, cluster 2 is low spending group and cluster 3 is medium spending group. In the case of K-Means clustering cluster 0 is the low spending group, cluster 1 is the high spending group and cluster 2 is the medium spending group.

Now let us check the different group of clusters and describe the profiles of each clusters.

High spending group (Cluster 1 in hierarchical and K-Means)

1. Convert their credit point to discount vouchers and they can use it during the purchase, since they are high spending group credit point also will be high.
2. Since they are high spending group try to provide offers on flight tickets and airport lounges, they may be frequent users of these things.
3. Since the probability of full payment and current balance of this group is high increase their credit limit.

4. Since their maximum pending in single shopping is high, try to provide more offers on luxury brands.

Medium spending group (Cluster 3 in hierarchical and cluster 2 in K-Means)

1. Since the probability of full payment is high for this group we can increase the credit limit of this group
2. Try to provide offers on daily use utensil, appliances, materials etc. and try to increase the purchase limit.
3. Offer premium cards and rewards points.

Low spending group (Cluster 2 in hierarchical and cluster 0 in K-Means)

1. Provide offers like increase in credit limit for a minimum amount spending of money within a time span.
2. Provide offers on payment in petroleum products, and waive off the transaction charges.
3. Keep regular contact with customer and provide reminder on repayment status.

PROBLEM 2 – 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 and compare the models' performances in train and test sets

INTRODUCTION

The purpose of this exercise is to build CART AND random forest to model to predict the insurance claim status based on the collected data from past few years, and to provide recommendation to management.

DATA DESCRIPTION

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)

SAMPLE OF THE DATASET

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA

Table 4 – Sample dataset

The dataset contains ten features.

EXPLORATORY DATA ANALYSIS

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64
2	Agency_Code	3000	object
3	Type	3000	object
4	Claimed	3000	object
5	Commision	3000	float64
6	Channel	3000	object
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	object
10	Destination	3000	object

Table 5 – Exploratory data analysis

There are total 3000 entries and 10 columns present in the dataset, and there are no null content present in the dataset.

DESCRIPTIVE DATA ANALYSIS

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000.0	NaN	NaN	NaN	38.091	10.463518	8.0	32.0	36.0	42.0	84.0
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000.0	NaN	NaN	NaN	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000.0	NaN	NaN	NaN	70.001333	134.053313	-1.0	11.0	26.5	63.0	4580.0
Sales	3000.0	NaN	NaN	NaN	60.249913	70.733954	0.0	20.0	33.0	69.0	539.0
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 6 – Descriptive data analysis

1. The data contains 3000 entries
2. The age of the passengers varies from 8 to 84 and the mean age is 38.
3. The data contains information from 4 agencies.
4. Most of the people from data did not claim the insurance.
5. Most of the people claimed through online.
6. There are total 5 types of plan available, and most people prefer customized plan.

Problem 2 – CART - RF

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 and compare the models' performances in train and test sets.

2.1 Read the data and do exploratory data analysis (4 pts). Describe the data briefly. Interpret the inferences for each (2 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.

The problem statement is to read the data and do the initial necessary steps and conduct exploratory data analysis on the dataset provided.

	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

Table 2.1.1 – Sample Dataset.

1. Age of insured (Age)
2. Code of tour firm (Agency_Code)
3. Type of tour insurance firms (Type)
4. Target: Claim Status (Claimed)
5. The commission received for tour insurance firm (Commission is in percentage of sales)
6. Distribution channel of tour insurance agencies (Channel)
7. Duration of the tour (Duration in days)
8. Amount worth of sales per customer in procuring tour insurance policies in rupees (in 100’s)
9. Name of the tour insurance products (Product)

10. Destination of the tour (Destination)

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64
2	Agency_Code	3000	object
3	Type	3000	object
4	Claimed	3000	object
5	Commision	3000	float64
6	Channel	3000	object
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	object
10	Destination	3000	object

Table 2.1.2 – Data info.

There are total 3000 entries in the data, and out of that 2 features are in int64 type, 2 features are in float 64, and 6 features are in object type data type.

Now let us check the null values in the dataset.

NO.	Column	Null content
1	Age	0
2	Agency_Code	0
3	Type	0
4	Claimed	0
5	Commision	0
6	Channel	0
7	Duration	0
8	Sales	0
9	Product Name	0
10	Destination	0

Table 2.1.3 – Null values.

There are no null value present in the dataset.

Next we can check the summary statistics of the given dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	3000.0	NaN	NaN	NaN	38.091	10.463518	8.0	32.0	36.0	42.0	84.0
Agency_Code	3000	4	EPX	1365	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Type	3000	2	Travel Agency	1837	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Claimed	3000	2	No	2076	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Commision	3000.0	NaN	NaN	NaN	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Channel	3000	2	Online	2954	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Duration	3000.0	NaN	NaN	NaN	70.001333	134.053313	-1.0	11.0	26.5	63.0	4580.0
Sales	3000.0	NaN	NaN	NaN	60.249913	70.733954	0.0	20.0	33.0	69.0	539.0
Product Name	3000	5	Customised Plan	1136	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Destination	3000	3	ASIA	2465	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2.1.4. – Summary statistics.

1. The data contains 3000 entries
2. The age of the passengers varies from 8 to 84 and the mean age is 38.
3. The data contains information from 4 agencies.
4. Most of the people from data did not claim the insurance.
5. Most of the people claimed through online.
6. There are total 5 types of plan available, and most people prefer customized plan.
7. Most of the peoples travel destination is Asia.

Next let us check the skewness of the data,

NO.	Column	Null content
1	Age	1.149713
2	Commission	3.148858
3	Duration	13.784681
4	Sales	2.381148

Table 2.1.5. – Skewness.

If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed. If the skewness is less than -1 or greater than 1, the data are highly skewed. Here the skewness value of data is more than 1, therefore we can conclude that the given data is highly skewed data.

Next we can check the outlier proportion of the given data,

NO.	Column	Number of outlier	Outlier proportion
1	Age	29	0.0096
2	Agency_Code	0	0
3	Type	0	0
4	Claimed	0	0
5	Commision	74	0.024
6	Channel	0	0
7	Duration	118	0.039
8	Sales	123	0.041
9	Product Name	0	0
10	Destination	0	0

Table 2.1.6. – Outlier proportion.

Outliers are present in all the int and float type data type. Most number of outlier present in sales and the outlier proportion is 0.041, and least number of outlier present in age column, and the outlier proportion is 0.0096.

Since the question did not mention about treating the outlier, we decided not to treat the outliers.

EXPLORATORY DATA ANALYSIS

Now let us do the exploratory data analysis of the given data.

UNIVARIATE ANALYSIS

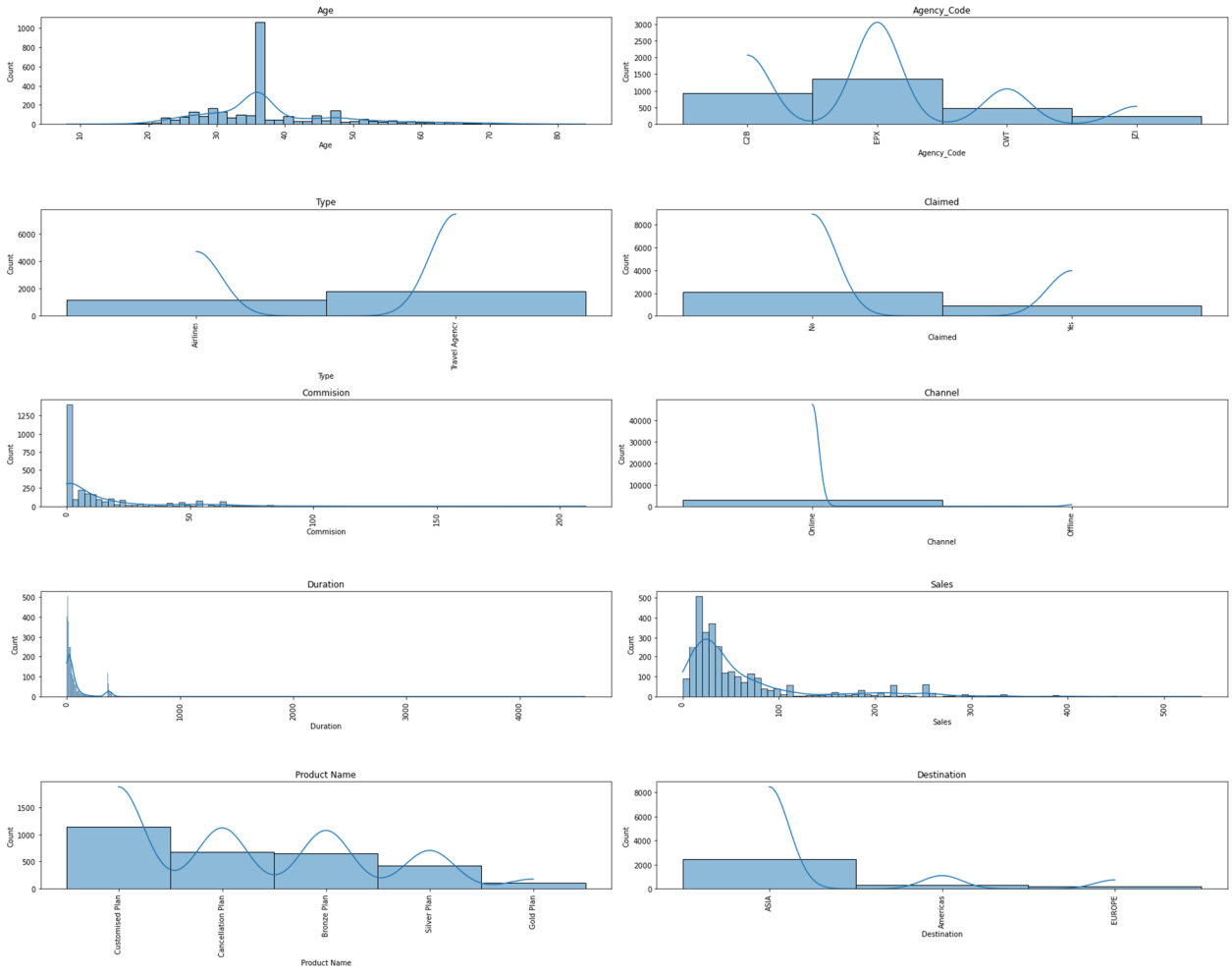

Fig 2.1.1 –Univariate analysis (HIst Plot)

The above shown is the hist plot of the different features of the dataset. The observations from the hist plot is given below.

1. Most of the travelers are in the age of range 30 to 40. And the passengers above age 80 is very less.
2. Most of the people travel through EPX agency and least people prefer JZI agency.
3. Most people prefer travel agency instead of airlines.
4. Most of the people did not claim their insurance.
5. In most of the cases the commission received for tour insurance firm is very less.

6. In the case of Distribution channel of tour insurance agencies, most people prefer online rather than offline.
7. Most of the travelers travel duration is very less.
8. Amount of sales is more in the range of 0 to 100.
9. Most favorite product among customers is customized plan and least favorite plan is gold plan.
10. Most of the peoples travel destination is Asia, and very few people travel to Europe.

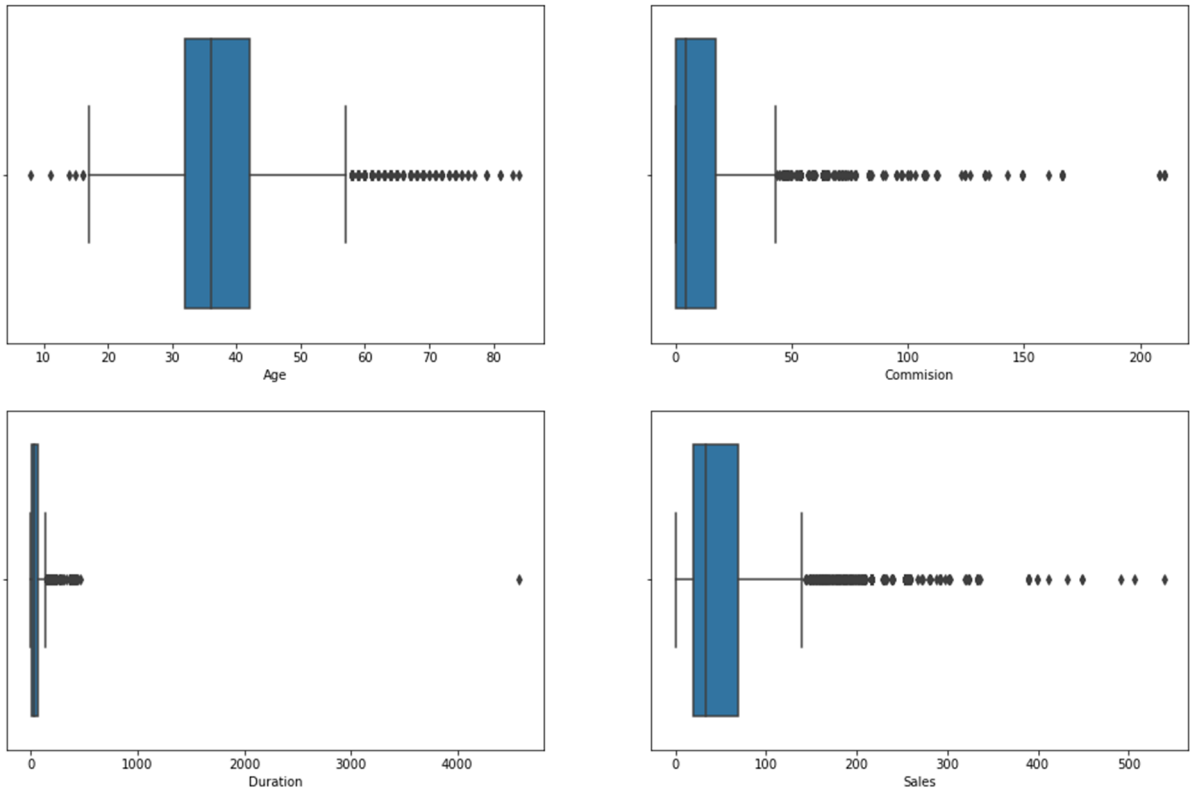

Fig 2.1.2 –Univariate analysis (Box Plot)

Now let us check the box plot of the numerical variable,

1. All the features are right skewed.
2. Most number of outliers present in sales column.
3. Duration have a negative number, may be a typing error.

BIVARIATE ANALYSIS

Now let us check the bivariate analysis of the categorical variables.

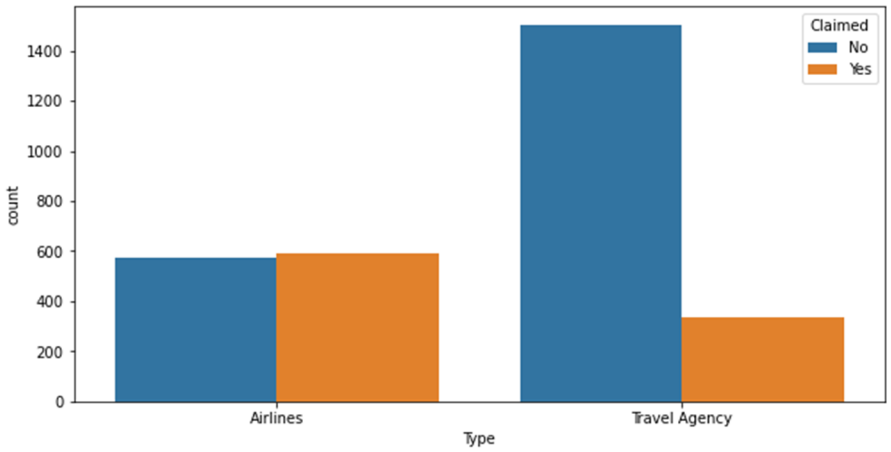

Fig 2.1.3 –Bivariate analysis (Type - Claimed)

The above plot shows the type of tour insurance firm and the insurance claimed. In the case of airlines the number of people claimed the insurance and not claimed the insurance is equal, whereas in the case of travel agency the number of people not claimed the insurance is very high compared to the number of people claimed the insurance.

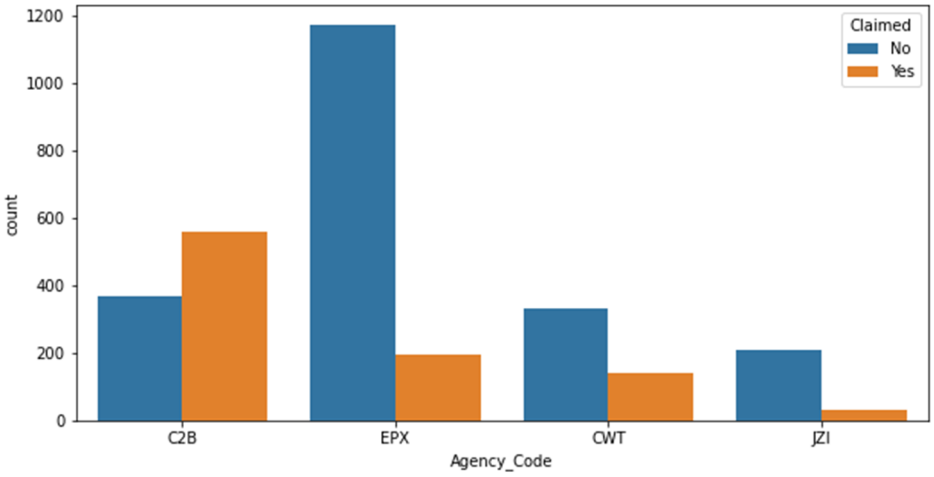

Fig 2.1.4 –Bivariate analysis (Agency code - Claimed)

The above plot shows the agency and the insurance claimed, in the case of C2B agency the number of people claimed the insurance is more compared to the number of people not claimed the insurance. In all other cases the number of people not claimed the insurance is more, and in the case of EPX agency the number of people not claimed the insurance is very high.

Next plot shows the relation between the product chosen and insurance claimed. In the case of gold plan and silver plan the number of people claimed the insurance is high compared to the people not claimed the insurance. In all other cases the number

of people not claimed the insurance is more, and in the case of customized plan it is very high.

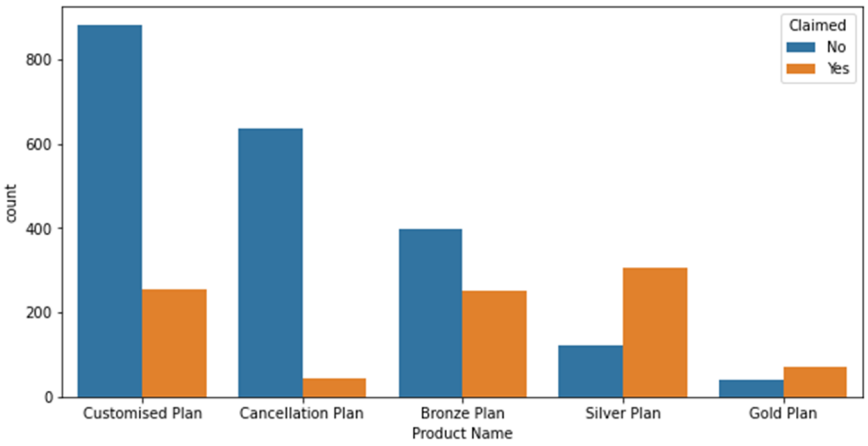

Fig 2.1.5 –Bivariate analysis (Product name - Claimed)

Next let us check the case of destination and the insurance claimed. As mentioned earlier most of the people travel to Asia, so in the case of insurance claimed also most number of people who did not claim the insurance are the travellers of Asia. In all the cases the number of people who did not claim the insurance is more than the number of people claim the insurance.

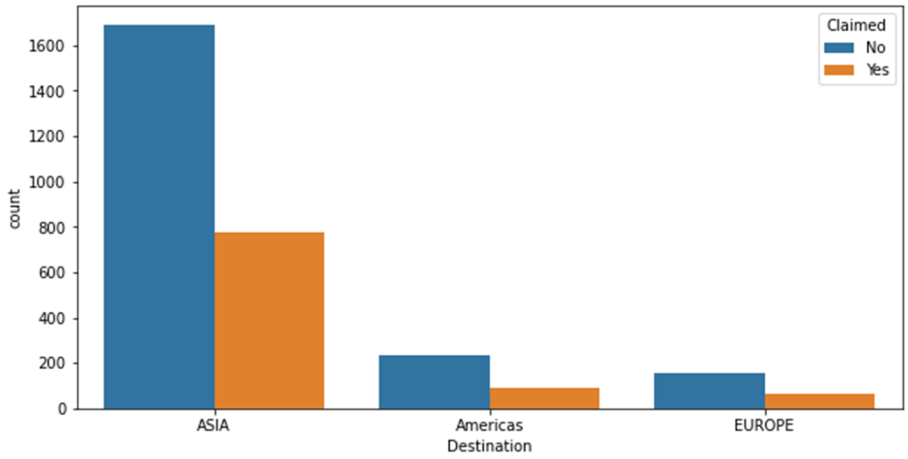

Fig 2.1.6 –Bivariate analysis (Destination - Claimed)

MULTIVARIATE ANALYSIS

Now let us check the pair plot and heat map of multivariate analysis.

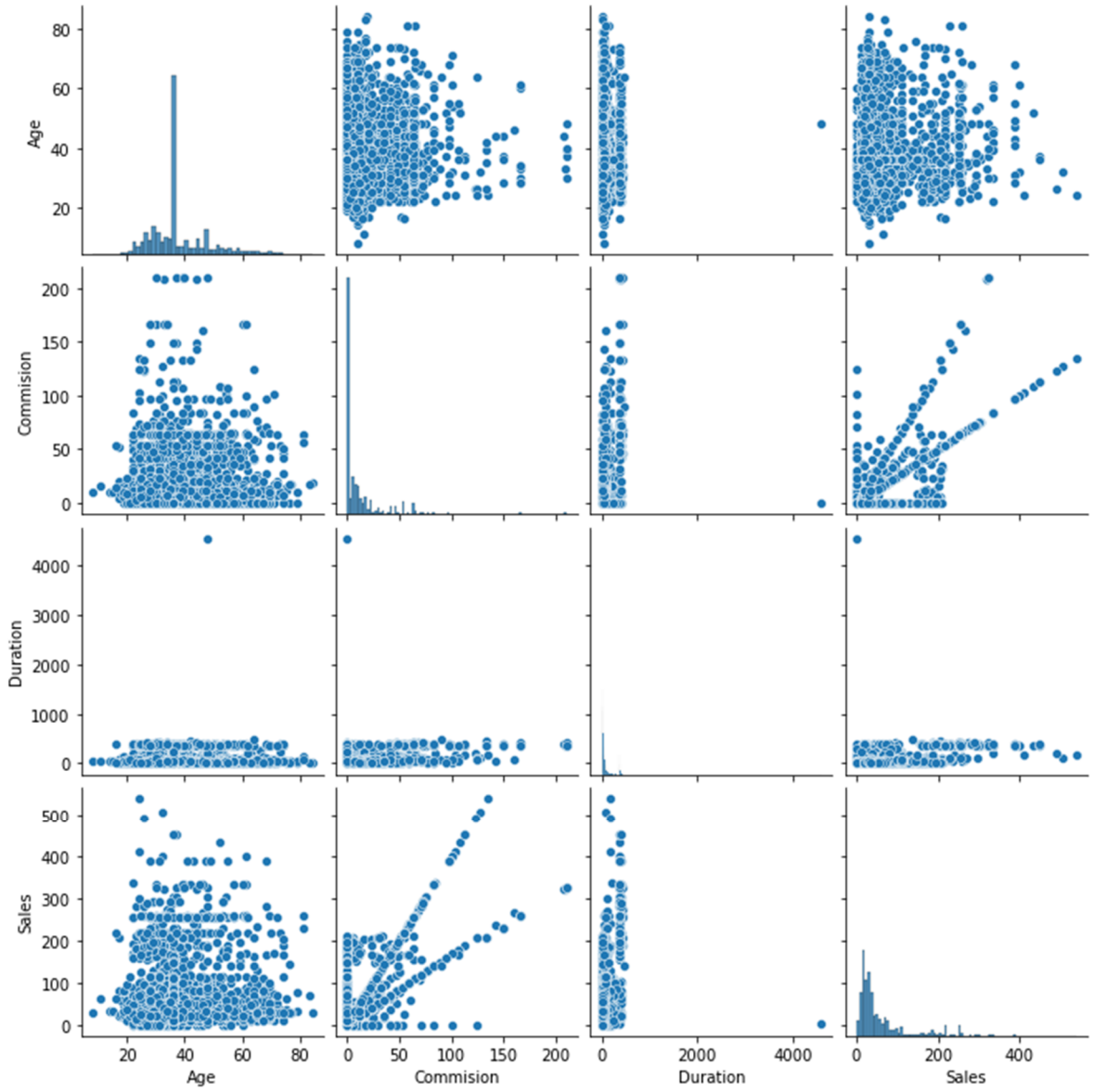

Fig 2.1.7 –Multivariate analysis (Pairplot)

The above figure shows the pairplot of the numerical features of the dataset. From the pair plot it is clear that the correlation between the different features in the dataset is very low.

The below shown plot shows the correlation heat map of the dataset. As mentioned in the case of pair plot from the heat map also the correlation value is very low between different features. The highest correlation is between sales and commission and it is 0.77.

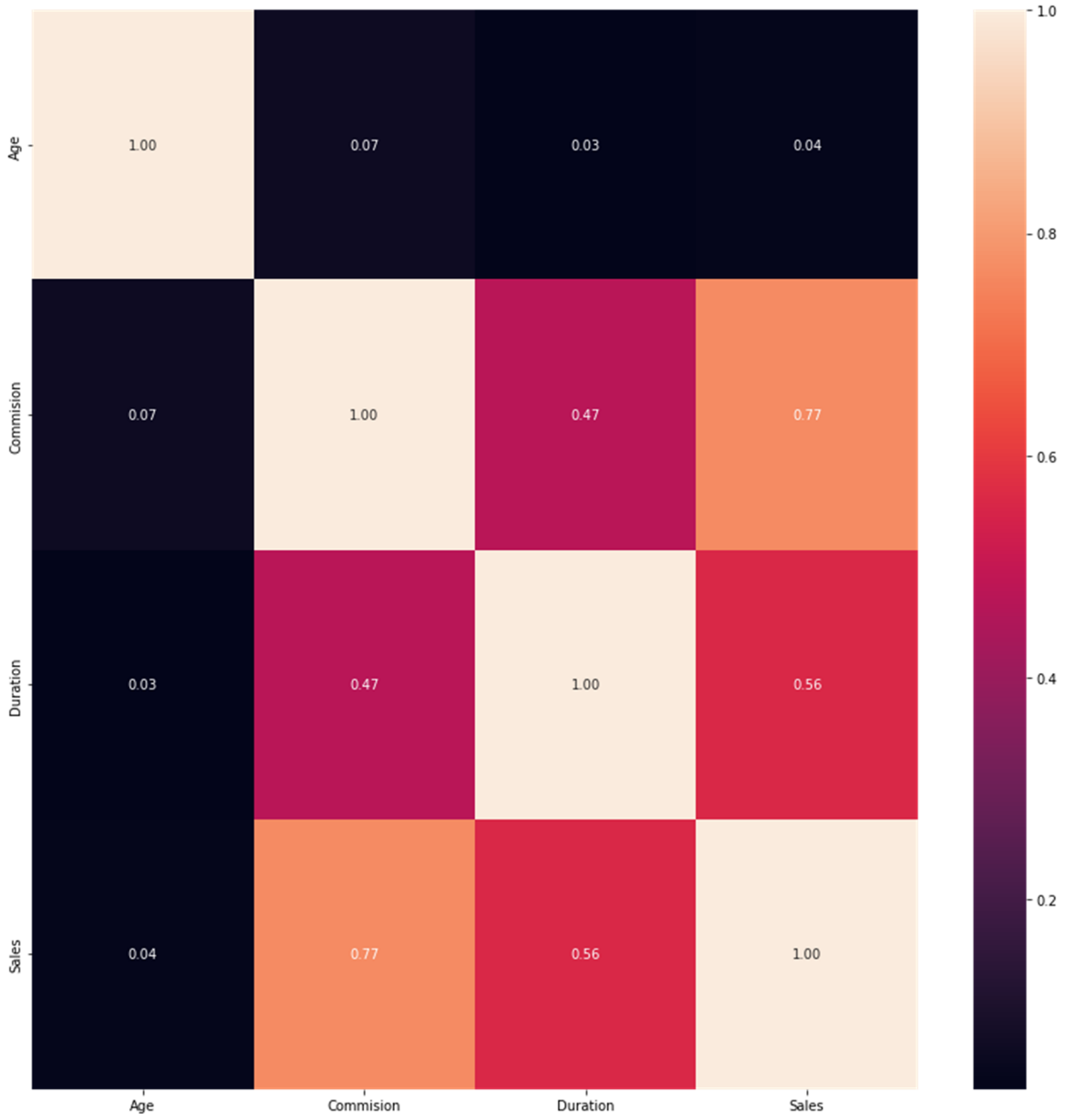

Fig 2.1.8 –Multivariate analysis (Heatmap)

2.2Data Split: Split the data into test and train(0.5 pts), build classification model CART (2.5 pts), Random Forest (2.5 pts). Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed. Any reasonable split is acceptable. Use of random state is mandatory. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Apply grid search for each model and make models on best_params. Feature importance for each model.

CART MODEL

The problem statement is to create decision tree and random forest model based on the given data to predict the insurance claim. Let us check the data info of the given dataset.

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64
2	Agency_Code	3000	object
3	Type	3000	object
4	Claimed	3000	object
5	Commision	3000	float64
6	Channel	3000	object
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	object
10	Destination	3000	object

Table 2.2.1 –Data info before conversion

For building decision tree and random forest model in python we have to ensure there are no object data type, there are only intiger data type for both dependent as well as independent data type. Here we have six features are in object data type. So we have to convert it into intiger data type using pd.categorical method and we have to extract code from that.

After the conversion of the dataset, let us check the latest data info,

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64

2	Agency_Code	3000	int8
3	Type	3000	int8
4	Claimed	3000	int8
5	Commision	3000	float64
6	Channel	3000	int8
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	int8
10	Destination	3000	int8

Table 2.2.2 –Data info after conversion

Now all the object data type converted to intiger data type.

Next we have to separate the data to independent variable and dependent variable. Here we have to predict the insurance claim status, so here claimed is the dependant variable and all others are independent variable.

Next step is splitting the data to training set and testing set. We can use train_test_split function from sklearn.model_selection for the splitting the data. Since the data is large enough we can go for a 70:30 ratio of train test split. The parameters we use for train test split are independent variable, dependent variable, splitting ratio ie, 70:30 in this case and we will consider random state as 1. The random state function is just like setting a seed value, this is to ensure uniformity when you running the same random generation function across multiple systems. After running the train test split we will get four outputs. Training independent variables, testing independent variables, training dependent variables and testing dependent variables. We will assign these values to separate variables.

The next step is to fit the model using the decision tree classifier and the criterion within the decision tree classifier as gini, and random state as 1 we are fitting the model. Gini is the measure of impurity of a dataset. We have to pass the independent as well as the dependent variables of the training data while building the model.

Let us check the plot of the decision tree we just build,

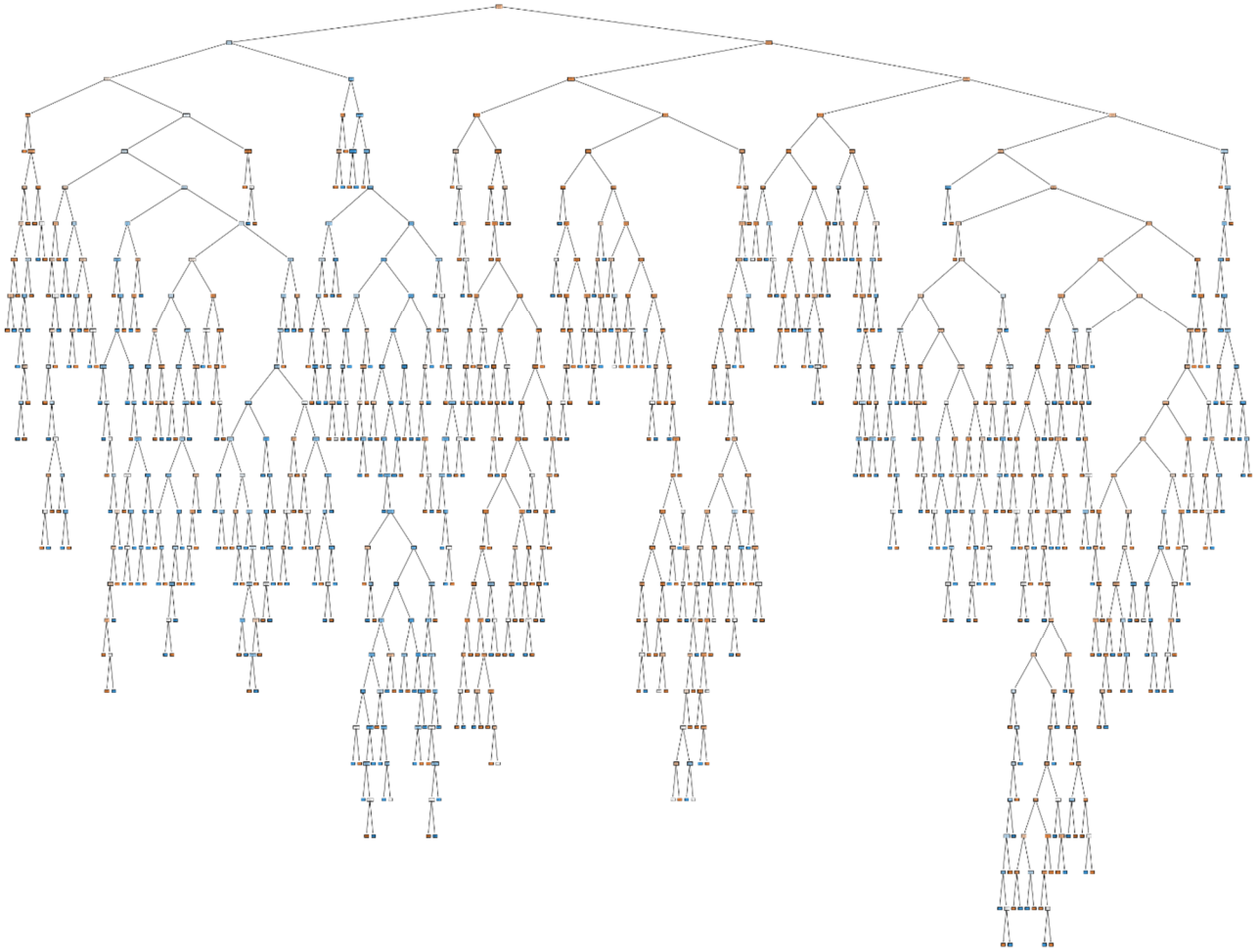

Fig 2.2.1 –Decision Tree Full

Since decision tree is very difficult to understand we will use pruning technique to minimalize the decision tree. We include certain other parameter also with gini for the pruning of the decision tree.

The other parameters we use are,

1. Max_depth - Maximum tree depth is a limit to stop further splitting of nodes when the specified tree depth has been reached during the building of the initial decision tree.
2. Min_sample_leaf - min_sample_leaf is the minimum number of samples required to be at a leaf node.
3. Min_sample_split - min_samples_split specifies the minimum number of samples required to split an internal node,

For finding the maximum depth of the decision tree we can use accuracy value of the decision tree at different depth and at what point accuracy score is maximum.

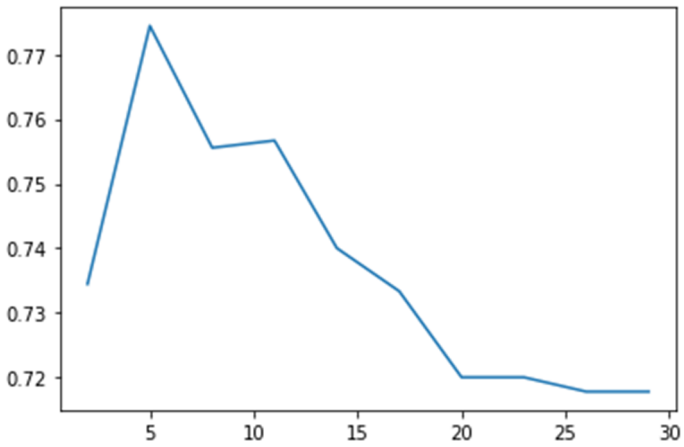

Fig 2.2.2 –accuracy plot

Above shown is the accuracy plot of the decision tree, at depth 5 accuracy value is maximum, so we can chose the depth of the decision tree as 5. Min sample leaf value is normally take as the 10% of the dataset and min sample split is three times the min sample leaf value.

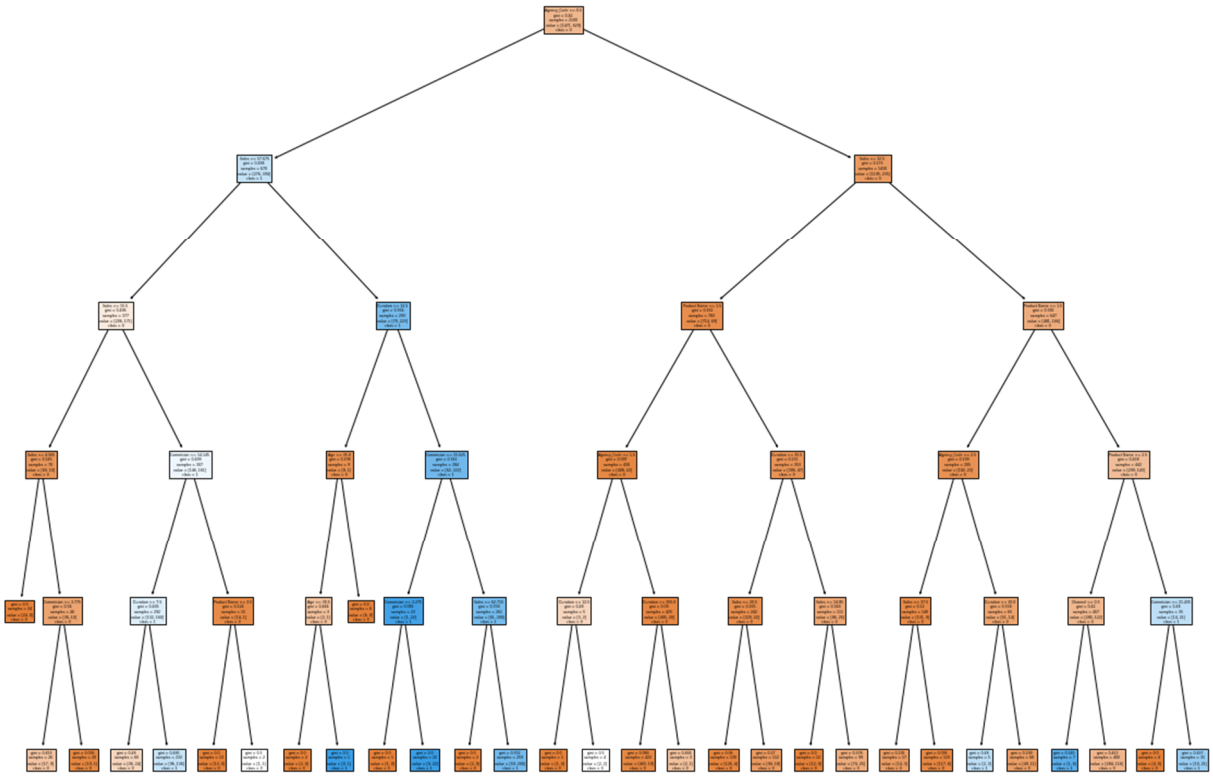

Fig 2.2.3 –Decision Tree after pruning

Let us do gridsearch on the model to find out the best parameters. The parameters we provided for gridsearch are,

'max_depth': [3, 4, 5, 6, 7],

'min_samples_leaf': [5, 15, 20, 25],

'min_samples_split': [15, 30, 45, 60]

And the output we got for the best parameters are {'max_depth': 4, 'min_samples_leaf': 5, 'min_samples_split': 45}
Accuracy score at depth 5 is 0.78 which is very good, so let us rebuild the decision tree at depth 5.

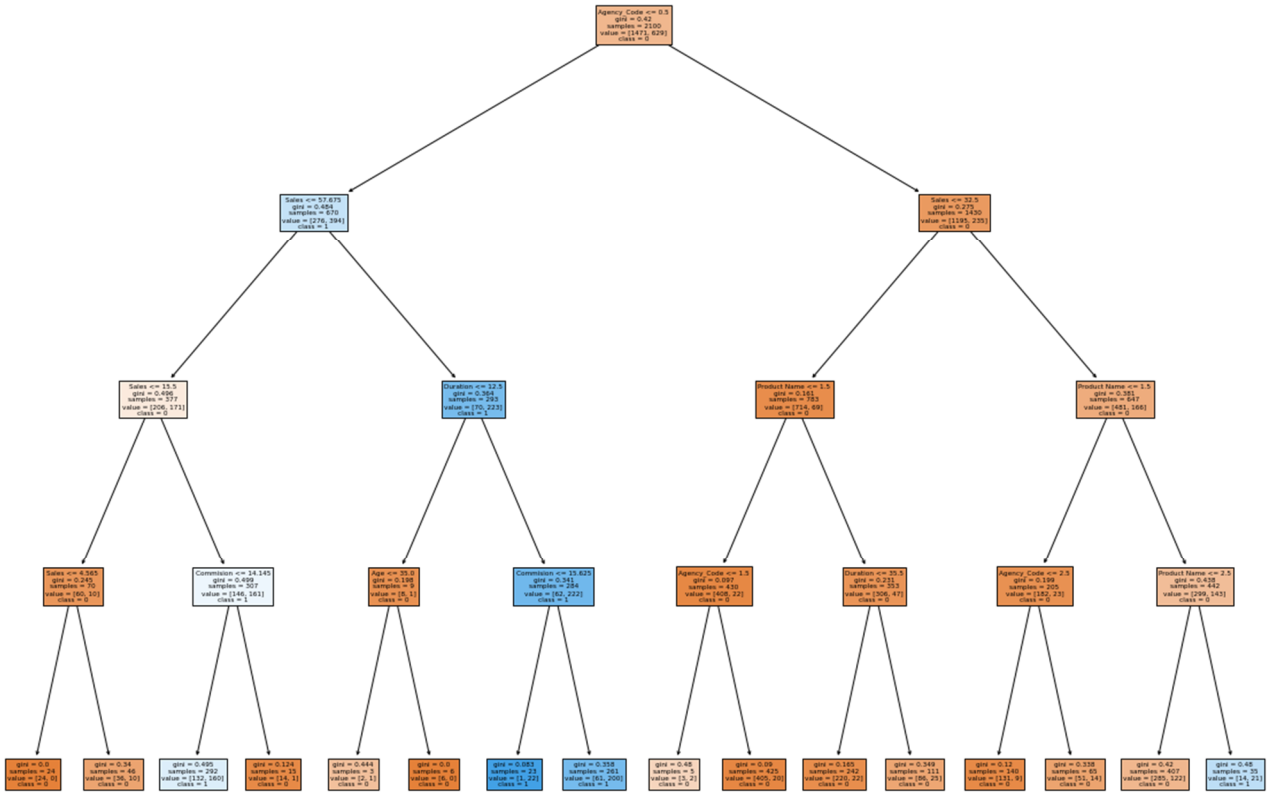

Fig 2.2.4 –Decision Tree rebuild.

Next let us check the feature importance of the decision tree model.

NO.	Column	Feature imp
1	Age	0.177894
2	Agency_Code	0.194770
3	Type	0.000383
4	Commision	0.095127

5	Channel	0.007262
6	Duration	0.262122
7	Sales	0.199864
8	Product Name	0.043258
9	Destination	0.019321

Table 2.2.3 –Feature importance

The table shows the feature importance value of the different independent variables. From the table it is clear that the most important feature among them all is duration of the travel and the least important feature is type of tour insurance firms.

RANDOM FOREST MODEL

The problem statement is to create decision tree and random forest model based on the given data to predict the insurance claim. Let us check the data info of the given dataset.

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64
2	Agency_Code	3000	object
3	Type	3000	object
4	Claimed	3000	object
5	Commision	3000	float64
6	Channel	3000	object
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	object
10	Destination	3000	object

Table 2.2.4 –Data info before conversion

For building decision tree and random forest model in python we have to ensure there are no object data type, there are only intiger data type for both dependent as well as independent data type. Here we have six features are in object data type. So we have to convert it into intiger data type using pd.categorical method and we have to extract code from that.

After the conversion of the dataset, let us check the latest data info,

NO.	Column	Non – Null content	Data Type
1	Age	3000	int64
2	Agency_Code	3000	int8

3	Type	3000	int8
4	Claimed	3000	int8
5	Commision	3000	float64
6	Channel	3000	int8
7	Duration	3000	int64
8	Sales	3000	float64
9	Product Name	3000	int8
10	Destination	3000	int8

Table 2.2.5 –Data info after conversion

Now all the object data type converted to intiger data type.

Next we have to separate the data to independent variable and dependent variable. Here we have to predict the insurance claim status, so here claimed is the dependant variable and all others are independent variable.

Next step is splitting the data to training set and testing set. We can use train_test_split function from sklearn.model_selection for the splitting the data. Since the data is large enough we can go for a 70:30 ratio of train test split. The parameters we use for train test split are independent variable, dependent variable, splitting ratio ie, 70:30 in this case and we will consider random state as 1. The random state function is just like setting a seed value, this is to ensure uniformity when you running the same random generation function across multiple systems. After running the train test split we will get four outputs. Training independent variables, testing independent variables, training dependent variables and testing dependent variables. We will assign these values to separate variables.

Let us create a random forest model with 1000 trees and let us check the out of bag score, the out of bag score we got is 0.755

Let us add more parameters to the random forest classifier and check the out of bag score. The parameters we add are,

1. n_estimators – number of decision trees to be created
2. Min_sample_leaf - min_sample_leaf is the minimum number of samples required to be at a leaf node.
3. Min_sample_split - min_samples_split specifies the minimum number of samples required to split an internal node.
4. Max_features - helps to find the number of features to take into account in order to make the best split.

We provided these values for the parameters,

n_estimators = 1000,

oob_score = True,

max_depth = 10,

max_features = 5,

min_samples_leaf = 50,

min_samples_split = 120,

And the out of bag score we got at this condition is 0.781.

For finding the maximum depth of the decision tree we can use accuracy value of the random forest at different depth and at what point accuracy score is maximum.

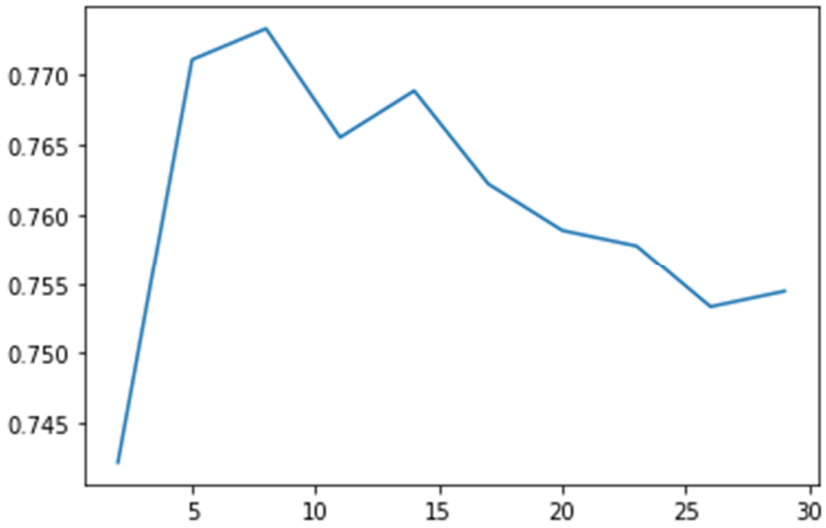

Fig 2.2.5 –accuracy plot

From the plot it is clear that the accuracy value is maximum at depth in between 7 and 10.

Let us do grid search on the model to get more clarity on the input parameters. The parameters we provided for grid search are,

'max_depth': [5,7,10],

'max_features': [4,6],

'min_samples_leaf': [5,10],

'min_samples_split': [50,100],

'n_estimators': [100,200,300]

And the output we got for the best parameters are 'max_depth': 7, 'max_features': 4, 'min_samples_leaf': 5, 'min_samples_split': 50, 'n_estimators': 100

By applying these parameters to the model the latest oob score is 0.793. So the oob score is increasing therefore the model is good.

Next let us check the feature importance of the random forest model,

NO.	Column	Feature imp
1	Age	0.056904
2	Agency_Code	0.302265
3	Type	0.045781
4	Commision	0.128050
5	Channel	0.003248
6	Duration	0.080293
7	Sales	0.176891
8	Product Name	0.196475
9	Destination	0.010092

Table 2.2.6 –Feature importance

The most important feature of the random forest is agency code and the least important feature is Distribution channel of tour insurance agencies.

2.3Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy (1 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC_AUC score for each model (2 pts), Make classification reports for each model. Write inferences on each model (2 pts). Calculate Train and Test Accuracies for each model. Comment on the validness of models (overfitting or underfitting) Build confusion matrix for each model. Comment on the positive class in hand. Must clearly show obs/pred in row/col Plot roc_curve for each model. Calculate roc_auc_score for each model. Comment on the above calculated scores and plots. Build classification reports for each model. Comment on f1 score, precision and recall, which one is important here.

The problem statement is to predict the accuracy score of train data and test data of decision tree as well as random forest. Let us first check the accuracy score of decision tree model.

Accuracy – How accurately / cleanly does the model classify the data points. Lesser the false predictions, more the accuracy

CART MODEL ACCURACY SCORE

Train Data

Let us first create the confusion matrix of the train data

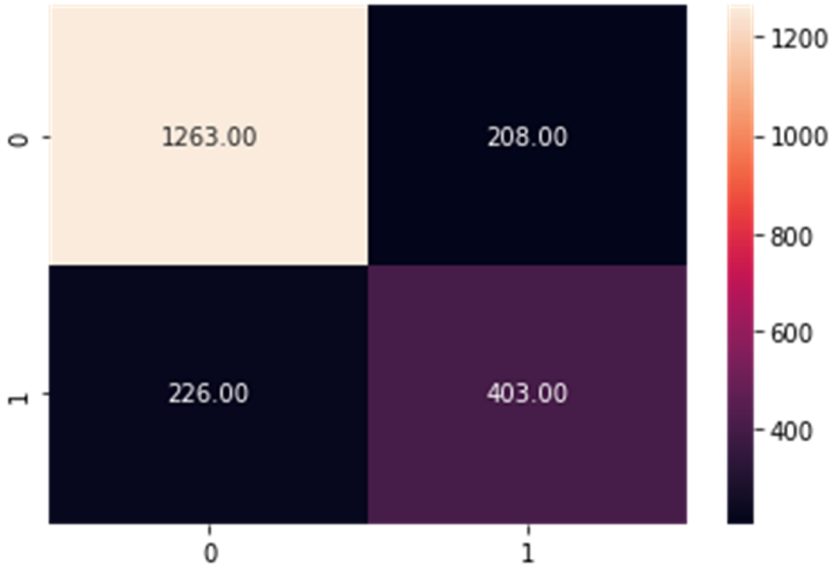

Fig 2.3.1 –CART confusion Matrix – Train data

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

TP : True Positives i.e. positive classes that are correctly predicted as positive.

FP: False Positives i.e negative classes that are falsely predicted as positive.

TN: True Negatives i.e. negative classes that are correctly predicted as negative.

FN: False Negatives i.e positive classes that are falsely predicted as negative.

$$\text{Accuracy} = \frac{(1263+403)}{(1263+403+208+226)}$$

$$= \frac{(1666)}{(2100)} = 0.7933$$

Test Data

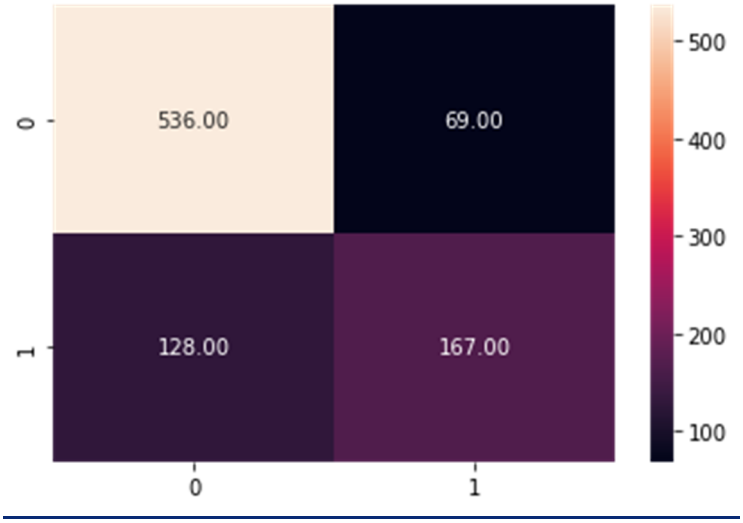

Fig 2.3.2 –CART confusion Matrix – Test data

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(536+167)}{(1263+403+208+226)}$$

$$= \frac{(703)}{(2100)} = 0.7811$$

RANDOM FOREST MODEL ACCURACY SCORE

Train Data

Let us first create the confusion matrix of the train data

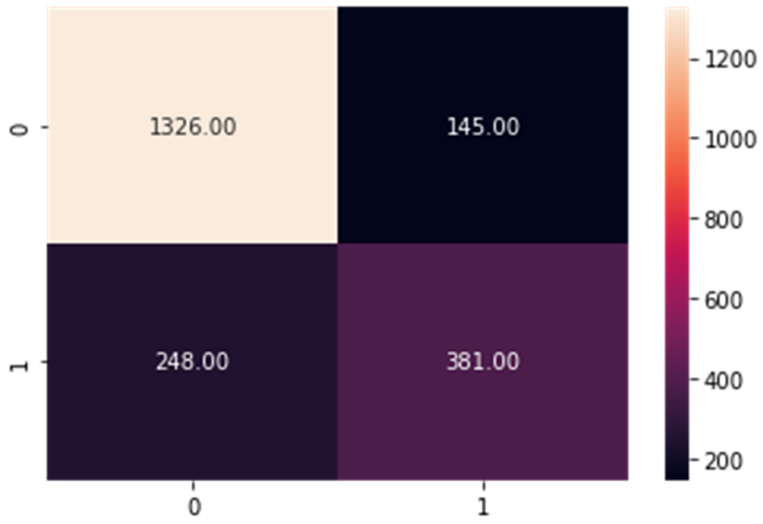

Fig 2.3.3 –RF confusion Matrix – Train data

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(1326+381)}{(1326+381+145+248)}$$

$$= \frac{(1707)}{(2100)} = 0.8128$$

Test Data

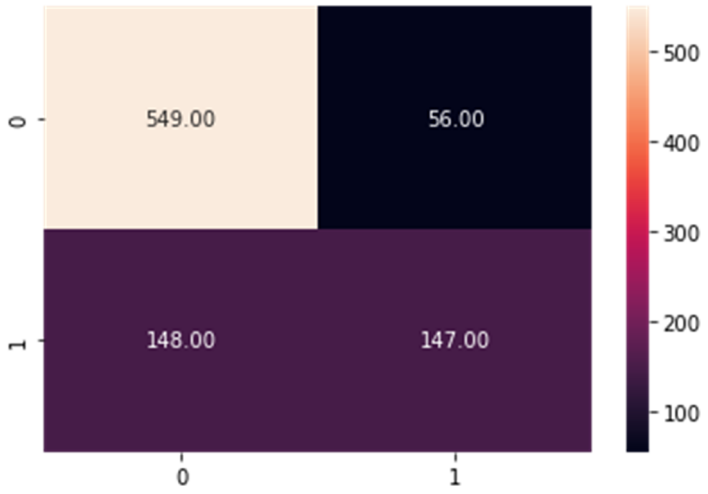

Fig 2.3.4 –RF confusion Matrix – Test data

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Accuracy} = \frac{(549+147)}{(549+147+56+148)}$$

$$= \frac{(696)}{(900)} = 0.7733$$

In both decision tree model and random forest model, the accuracy value of the train data and is more than the accuracy value of the test data. This indicates the over fitting of the model. Since only 80% accuracy is only showing the model in both decision tree and random forest we should consider increase the amount of data sample size.

Let us consider the above confusion matrix for CART test data,

NO.	Predicted value	
Actual	0(Predicted negative)	1(Predicted positive)
0(Actual negative)	536	69
1(Actual positive)	128	167

Table 2.3.1 –CART Prediction Test data

NO.	Predicted value	
Actual	0(Predicted negative)	1(Predicted positive)
0(Actual negative)	549	56
1(Actual positive)	148	147

Table 2.3.2 –RF Prediction Test data

In both decision tree and random forest model, the value of true negative is more compared to the value of true positive. Let us calculate the sensitivity of the model to determine how apt the model is to detecting events in the positive class.

Sensitivity of CART

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)}$$

$$\text{Sensitivity} = \frac{(167)}{(167+128)}$$

$$= \frac{(167)}{(295)} = 0.566$$

Sensitivity of RF

$$\text{Sensitivity} = \frac{(TP)}{(TP+FN)}$$

$$\text{Sensitivity} = \frac{(147)}{(147+148)}$$

$$= \frac{(147)}{(295)} = 0.49$$

The value of sensitivity is very low, so we can conclude that model has less accuracy predicting true positive values.

ROC – AUC CURVE OF CART

Now let us check the ROC – AUC curve of testing and training data of CART model,

Train Data

We use the probability value of positive outcomes for the construction of ROC curve as well as the AUC score, We use ROC – AUC curve to compare the model. The probability value of train dataset is shown below,

	0	1
0	0.909091	0.090909
1	0.700246	0.299754
2	0.233716	0.766284
3	0.452055	0.547945
4	0.400000	0.600000
5	0.452055	0.547945
6	0.700246	0.299754
7	0.935714	0.064286
8	0.952941	0.047059
9	0.700246	0.299754

Table 2.3.3 –Probability value of predicted Train dataset

By using the positive outcomes from the probability of the predicted training dataset we will calculate the area under the curve (AUC) value of the train dataset, The AUC value we got for training dataset is,

AUC for Training data = 0.833

Next let us construct the ROC (Receiver Operating Characteristics) curve for the training dataset of the CART model, we will use roc_curve parameter from sklearn.metrics for the construction of the ROC curve, we will use dependant train

data and probability of the predicted training data as the parameters for the ROC curve. The constructed ROC curve is shown below,

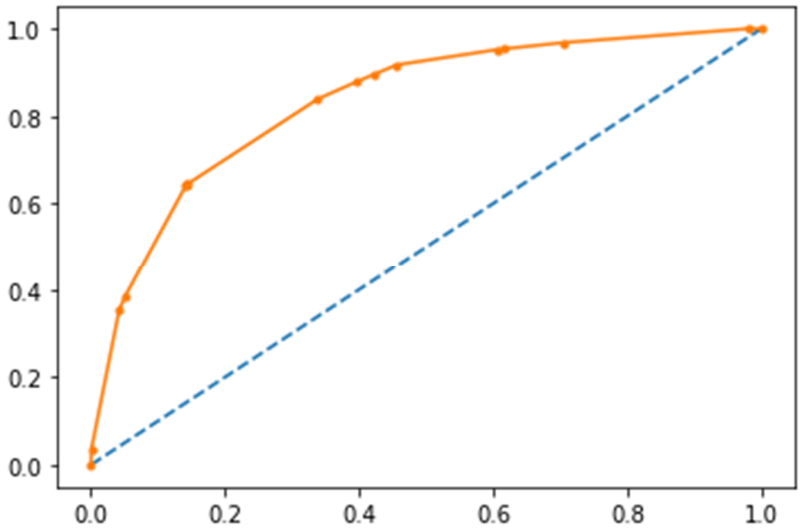

Fig 2.3.5 –CART ROC Curve – Train data

Test Data

For calculating the AUC score and constructing the ROC curve for the test data of the CART model, we will use the same procedure we did for the train data,

	0	1
0	0.935714	0.064286
1	0.452055	0.547945
2	0.452055	0.547945
3	0.233716	0.766284
4	0.952941	0.047059
5	0.700246	0.299754
6	0.952941	0.047059
7	0.700246	0.299754
8	0.700246	0.299754
9	0.452055	0.547945

Table 2.3.4 –Probability value of predicted Test dataset

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Testing data = 0.798

Next using dependant test data and probability of the predicted test data, we will create the ROC curve for the test data and it is shown below,

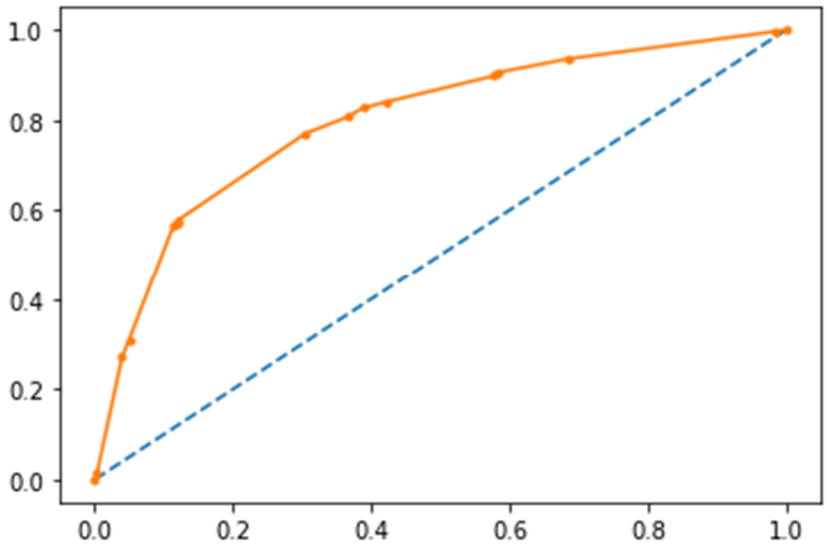

Fig 2.3.6 –CART ROC Curve – Test data

We got the AUC score of the test and train data in the range of 0.7 to 0.8 which is acceptable, as expected the AUC score of the testing data is less than the AUC score of the training data. The same is visible in the case of ROC curve too.

ROC – AUC CURVE OF RANDOM FOREST

Next let us calculate the AUC score and construct the ROC curve of random forest model for test as well as train data,

Train Data

As mentioned in the case of CART model, similar steps will follow in the case of random forest model too,

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Training data = 0.864

Next using dependant train data and probability of the predicted train data, we will create the ROC curve for the test data and it is shown below,

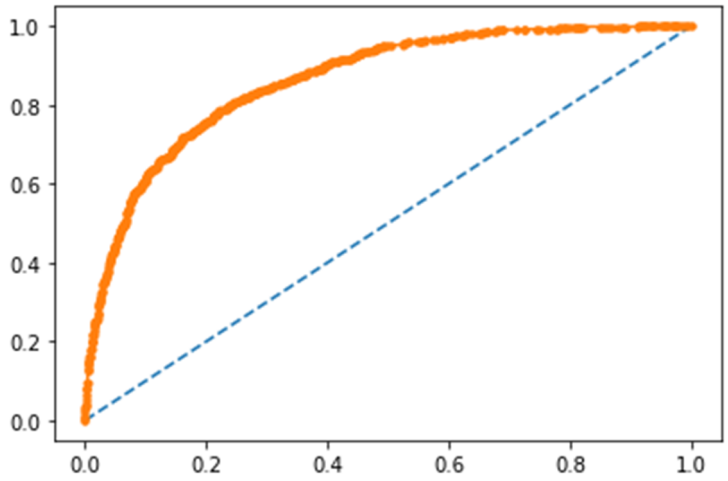

Fig 2.3.7 –RF ROC Curve – Train data

Test Data

By using the positive predicted values from the predicated probability test data set we will calculate the value of AUC score,

AUC for Test data = 0.823

Next using dependant test data and probability of the predicted test data, we will create the ROC curve for the test data and it is shown below,

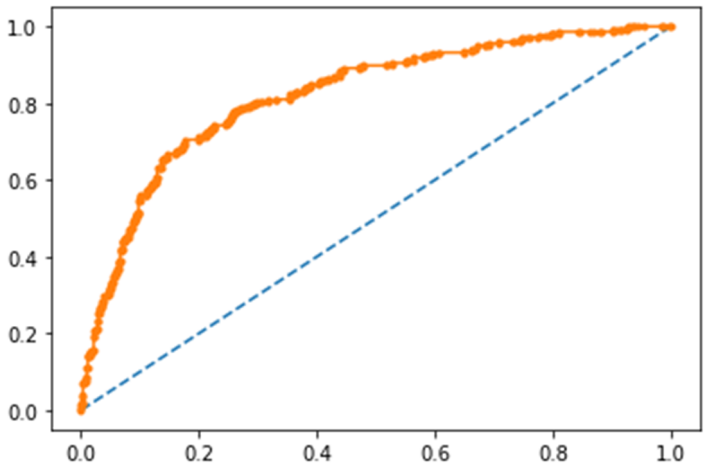

Fig 2.3.8 –RF ROC Curve – Test data

As similar to CART model the AUC score of the Random forest model also is high for the train data than the test data. We got the AUC score in the range 0.8 to 0.9 considered excellent. In the case of ROC curve also same scenario is visible, area under the curve is more for the train data than the test data.

CLASSIFICATION REPORT FOR CART MODEL

A classification report is a performance evaluation metric in machine learning. It is used to show the precision, recall, F1 Score, and support of your trained classification model. Let understand each parameters,

Precision - Precision is defined as the ratio of true positives to the sum of true and false positives.

$$\text{Precision} = \frac{(TP)}{(TP+FP)}$$

Recall - Recall is defined as the ratio of true positives to the sum of true positives and false negatives.

$$\text{Recall/Sensitivity} = \frac{(TP)}{(TP+FN)}$$

TP : True Positives i.e. positive classes that are correctly predicted as positive.

FP: False Positives i.e negative classes that are falsely predicted as positive.

TN: True Negatives i.e. negative classes that are correctly predicted as negative.

FN: False Negatives i.e positive classes that are falsely predicted as negative.

F1-score - The F1 is the weighted harmonic mean of precision and recall. The closer the value of the F1 score is to 1.0, the better the expected performance of the model is.

Support - Support is the number of actual occurrences of the class in the dataset. It doesn't vary between models, it just diagnoses the performance evaluation process.

Let us check the classification report of train and test data of CART model,

Train data

The classification report for train data of the CART model is shown below,

	precision	recall	f1-score	support
0	0.85	0.86	0.85	1471
1	0.66	0.64	0.65	629
accuracy			0.79	2100
macro avg	0.75	0.75	0.75	2100
weighted avg	0.79	0.79	0.79	2100

Table 2.3.5 –CART – Classification Report – Train Data

Test data

The classification report for test data of the CART model is shown below,

	precision	recall	f1-score	support
0	0.81	0.89	0.84	605
1	0.71	0.57	0.63	295
accuracy			0.78	900
macro avg	0.76	0.73	0.74	900
weighted avg	0.77	0.78	0.77	900

Table 2.3.6 –CART – Classification Report – Test Data

CLASSIFICATION REPORT FOR RANDOM FOREST MODEL

Similar to CART model the classification report of train and test data random forest model is shown below,

Train data

The classification report for train data of the random forest model is shown below,

	precision	recall	f1-score	support
0	0.84	0.90	0.87	1471
1	0.72	0.61	0.66	629
accuracy			0.81	2100
macro avg	0.78	0.75	0.77	2100
weighted avg	0.81	0.81	0.81	2100

Table 2.3.7 –RF – Classification Report – Train Data

Test data

The classification report for test data of the random forest model is shown below,

	precision	recall	f1-score	support
0	0.79	0.91	0.84	605
1	0.72	0.50	0.59	295
accuracy			0.77	900
macro avg	0.76	0.70	0.72	900
weighted avg	0.77	0.77	0.76	900

Table 2.3.8 –RF – Classification Report – Test Data

2.4 Final Model - Compare all models on the basis of the performance metrics in a structured tabular manner (2.5 pts). Describe on which model is best/optimized (1.5 pts). A table containing all the values of accuracies, precision, recall, auc_roc_score, f1 score. Comparison between the different models(final) on the basis of above table values. After comparison which model suits the best for the problem in hand on the basis of different measures. Comment on the final model.

Now let us compare all the performance metrics values of train data and test data of decision tree and random forest, consider below table,

	CART TRAIN	CART TEST	RANDOM FOREST TRAIN	RANDOM FOREST TEST
ACCURACY	0.7933	0.7811	0.8128	0.7733
AUC	0.833	0.798	0.864	0.823
PRECISION	0.66	0.71	0.72	0.72
RECALL	0.64	0.57	0.61	0.50
F1-SCORE	0.65	0.63	0.66	0.59

Table 2.4.1 –Performance metrics comparison

The table shows the comparison of test and train data of decision tree and random forest based on the performance metrics.

Let us compare the ROC curve of test and train data of CART and RF as well,

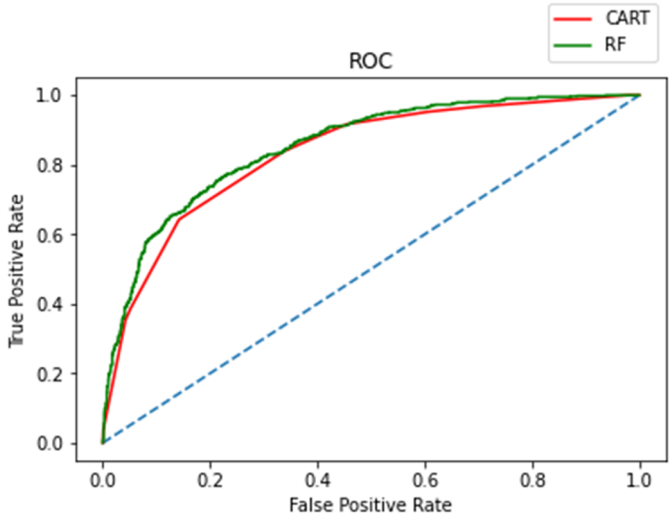

Fig 2.4.1 –CART – RF – ROC Curve – Train data

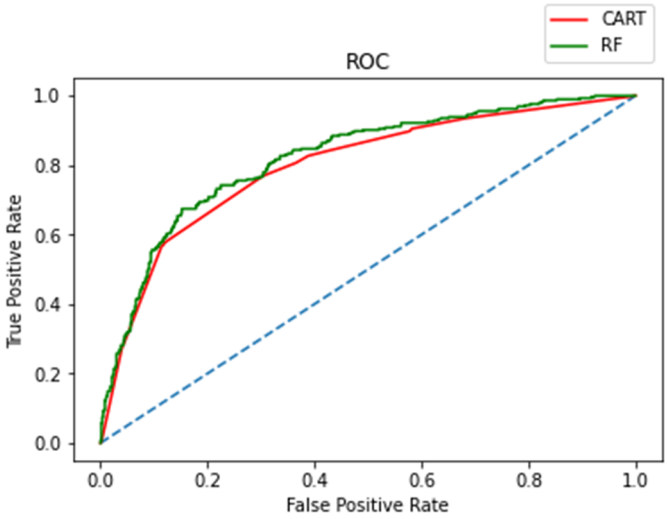

Fig 2.4.2 –CART – RF – ROC Curve – Test data

From the above table we can clearly understand that the value of all the model performance measures are higher for the random forest than the CART model. In the case of comparison ROC curve, area under the curve of random forest for test and train data is more than the area under the curve of test and train data of the CART model.

From the above observations we can confirm that the best model is random forest model. But when we check the classification report we can see that the precision rate of prediction of 0 is higher than the precision rate of prediction of 1. That is this model is more accurate in predicting not claimed status of the insurance rather than the claimed status of the insurance. From the descriptive data analysis 2/3 of the data is contains the information of people who did not claim the insurance. So, we should consider collecting more past data for the analysis for the better performance of the model.

2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

1. As per the problem statement that the insurance firm facing a higher claim frequency and the random forest model we build have more accuracy in predicting people who did not claim the insurance. So it will help them.
2. The agency code have good feature importance in random forest model, to try to tie up with more number of travel agencies to widen the business.
3. Maintain good customer relationship with those customers who did not claim the insurance, their repeated visit will increase the profit of the firm.
4. Provided training for JZI agency to increase the business and maintain a very good relation with EPX agency since most of the business came from them.
5. Introduce more number of plans and increase the features of the gold plan and increase the sale, also find out why claim ratio is more for only gold plan.
6. Most of the business came from travel agency but the insurance claim rate is more for airlines, find out the reason.

THE END

CLICK HERE TO GO TO CONTENTS