DATA MINING ASSIGNMENT

BUSINESS REPORT

JASPER SHELDON M

Contents

[Problem 1: Clustering 3](#_Toc81172639)

[1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 3](#_Toc81172640)

[1.2 Do you think scaling is necessary for clustering in this case? Justify 8](#_Toc81172641)

[1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them 9](#_Toc81172642)

[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. 12](#_Toc81172643)

[1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. 15](#_Toc81172644)

[Problem 2: CART-RF-ANN 16](#_Toc81172645)

[2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). 16](#_Toc81172646)

[2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network 25](#_Toc81172647)

[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. 26](#_Toc81172648)

[2.4 Final Model: Compare all the models and write an inference which model is best/optimized. 32](#_Toc81172649)

[2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations 33](#_Toc81172650)

|  |  |  |
| --- | --- | --- |
| List of Figures |  |  |
| **S.No** | **Name** | **Pg.No** |
| 1 | Figure 1 Duplicates | 5 |
| 2 | Figure 2 .Histogram and Boxplot | 6 |
| 3 | Figure 3. Correlation Plot | 7 |
| 4 | Figure 4 Pair Plot | 8 |
| 5 | Figure 5 Dendogram | 9 |
| 6 | Figure 6. Boxplot of clusters based on spending | 10 |
| 7 | Figure 7 Pairplot of the clusters | 11 |
| 8 | Figure 8 Elbow plot | 13 |
| 9 | Figure 9 Silhouette score | 14 |
| 10 | Figure 10 Boxplot of numerical columns | 18 |
| 11 | Figure 11 Unique values of categorical columns | 19 |
| 12 | Figure 12 Boxplot on age | 19 |
| 13 | Figure 13 Agency code vs Claim | 20 |
| 14 | Figure 14 Type vs Claim | 20 |
| 15 | Figure 15 Commision vs type | 21 |
| 16 | Figure 16 Commission by class | 21 |
| 17 | Figure 17 Commission vs. Agency | 22 |
| 18 | Figure 18 calims in each channel | 22 |
| 19 | Figure 19 Duration vs. claim | 23 |
| 20 | Figure 20 Plan type vs claim | 23 |
| 21 | Figure 21 Destination vs claim | 24 |
| 22 | Fig 22 | 24 |
| 23 | Figure 23 Roc curve for decison tree | 26 |
| 24 | Figure 24 Roc curve test dataset | 27 |
| 25 | Figure 25 Roc curve RF for train dataset | 28 |
| 26 | Figure 26 Roc curve RF for test set | 29 |
| 27 | Figure 27 Roc curve Ann Train data | 30 |
| 28 | Figure 28 Roc curve ann test data | 31 |

|  |  |  |
| --- | --- | --- |
| List of Tables |  |  |
| **S.No** | **Name** | **Pg.No** |
| 1 | Table 1.Sample data | 4 |
| 2 | Table 2.Summary of the data | 4 |
| 3 | Table 3. Data Information | 5 |
| 4 | Table 4. Skewness of the data | 5 |
| 5 | Table 5 Scaled Data | 9 |
| 6 | Table 6 Wss values | 13 |
| 7 | Table 7 Cluster Profiles | 15 |
| 8 | Table 8 Sample of the data | 17 |
| 9 | Table 9 Information of data | 17 |
| 10 | Table 10 Summary of the data | 18 |
| 11 | Table 1 Performance metrics | 34 |

# 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)**.

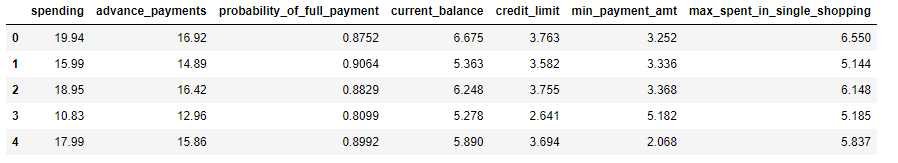


Table 1.Sample data

From the table above we can see that all the variables are continuous and no categorical variable is present. Along with that we can see the column **PROBABILITY OF FULL PAYMENT** is in 1/10th range.

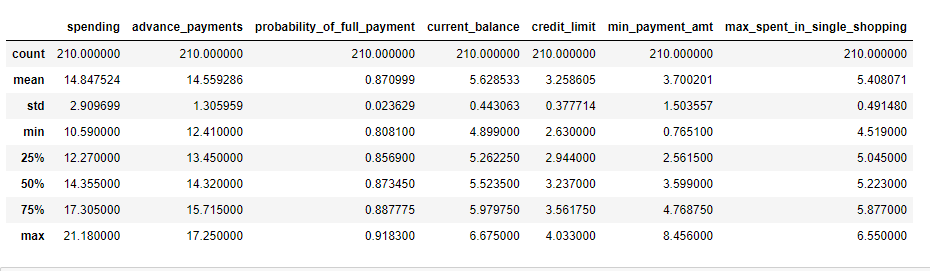


Table 2.Summary of the data

**SPENDING** and **ADVANCE PAYMENTS** are in the same scale. The standard deviation of **PROBABILITY OF FULL PAYMENTS** is very less compared to the other columns. Comparing the mean, standard deviation and the quantiles we don’t find any flawed data points.

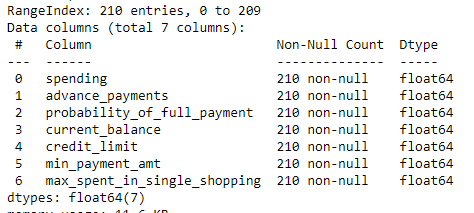


Table 3. Data Information

From the table above we can see there are total of 210 customers and 7 columns. All the columns indicate they have 210 non null value so it means there are no missing values present in the dataset also all 7 columns are continuous and of float type.

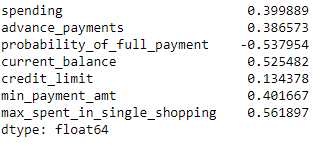


Table 4. Skewness of the data

From the table we can see only **PROBABILITY OF FULL PAYMENT** negatively (left) skewed and all the other columns are positively (right) skewed, except for **CREDIT LIMIT** all the other columns have a significant skewness in them.

One more key note there is no duplicates in the dataset we can infer this from the below output.

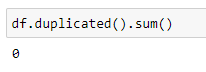
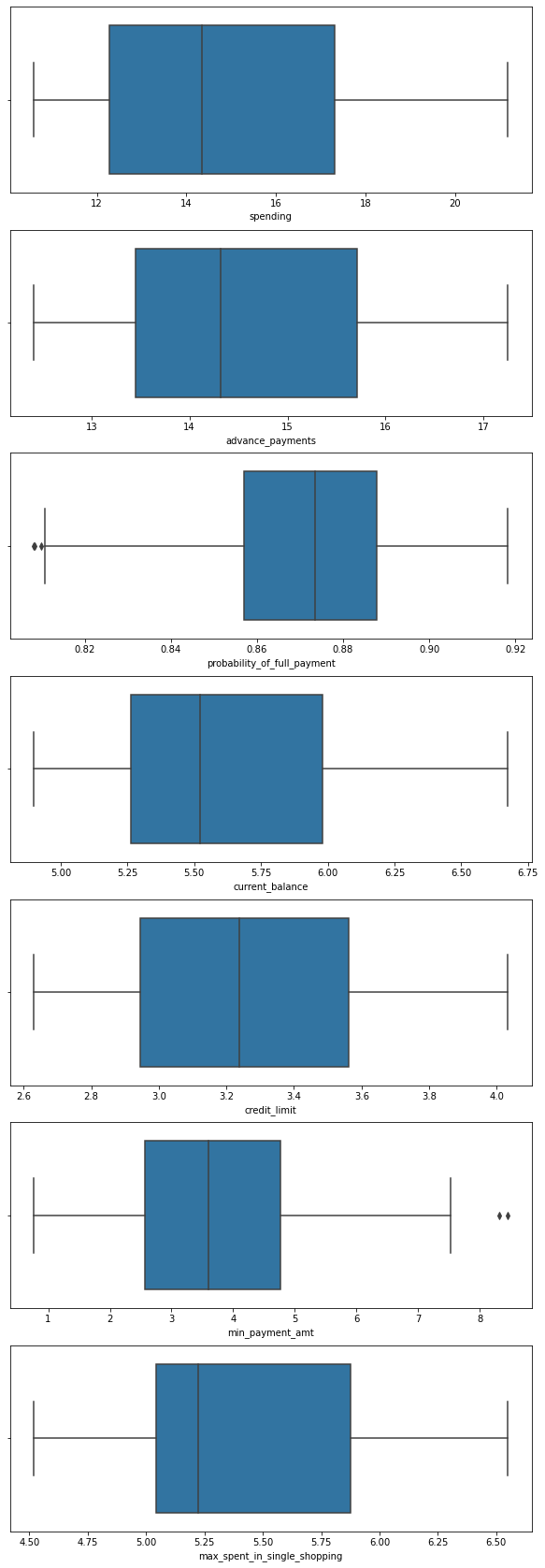
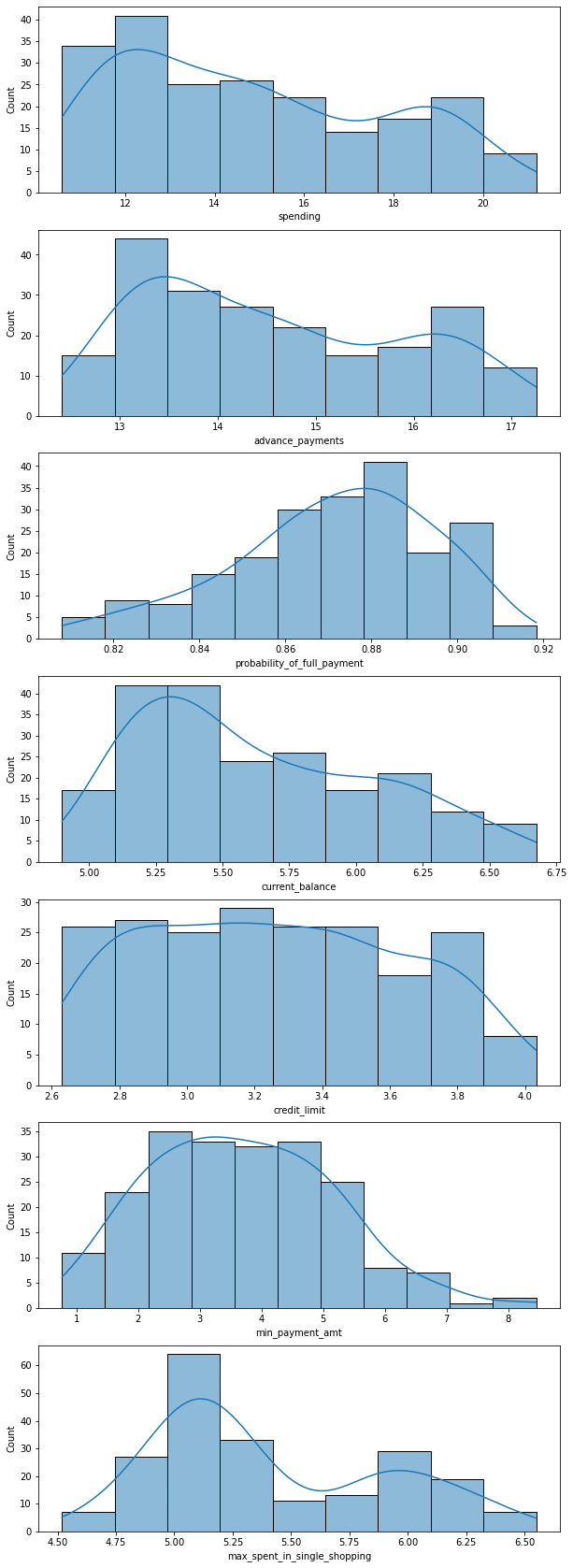


Figure 1 Duplicates

**Figure 2 .Histogram and Boxplot**

From the figure we can see that **PROBABILITY OF FULL PAYMENTS** and **MIN PAYMENT AMT** almost follows a normal distribution. There are totally 4 outliers present in the dataset 2 in **PROBABILITY OF FULL PAYMENTS** and 2 in **MIN PAYMENT AMT.** Since we are going to do clustering we’ll need to handle these outliers.

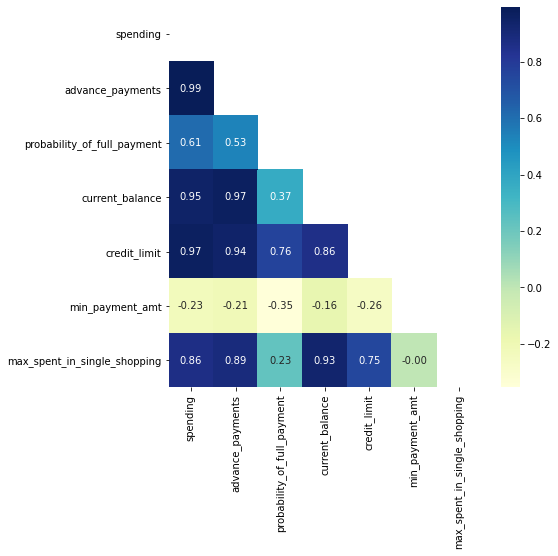


Figure 3. Correlation Plot

From the above figure we can see **MIN PAYMENT AMT** is negatively correlated with all the columns and there are high positive correlation existing among the variables.

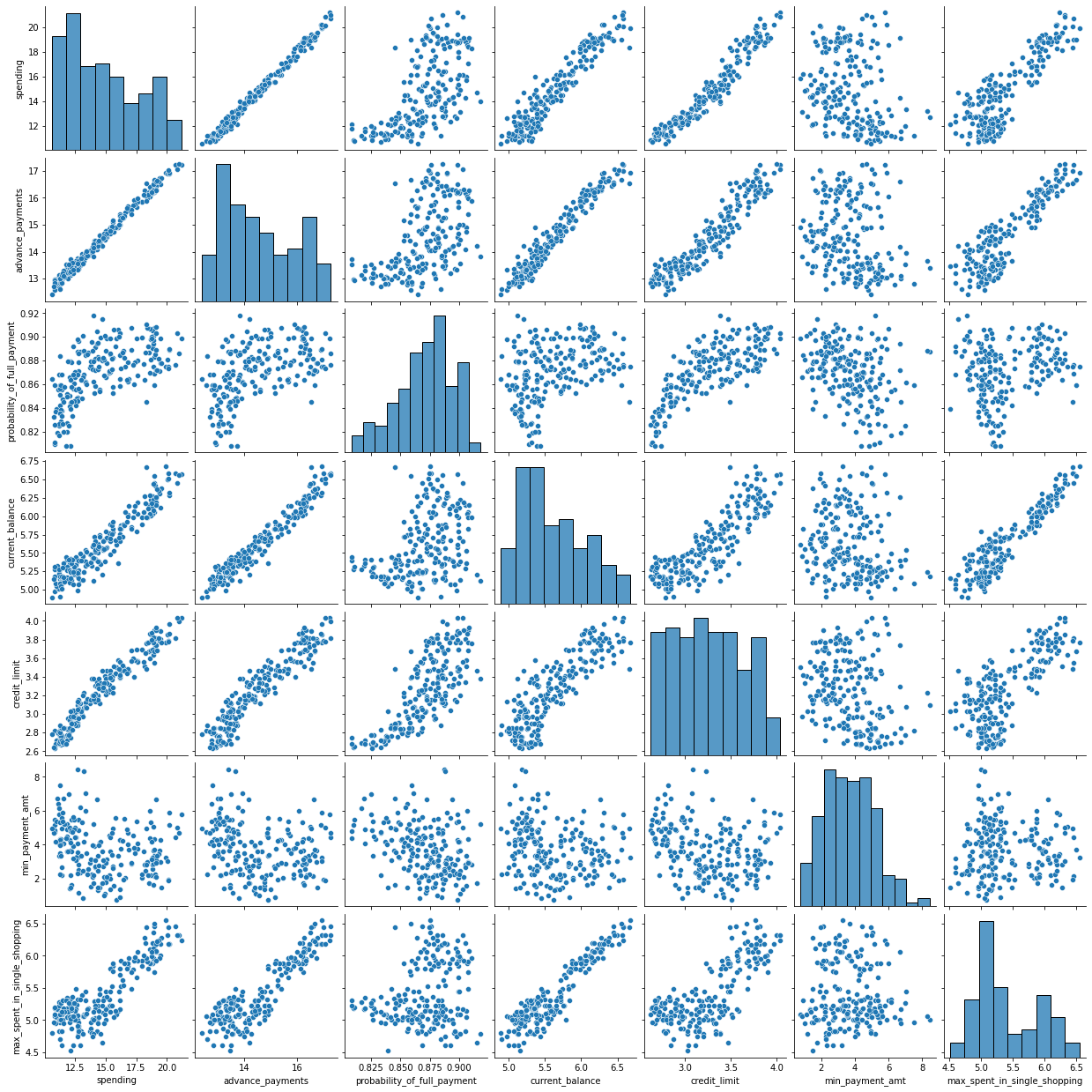


Figure 4 Pair Plot

We can see most of the variables has positive linear relationships among them.

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

Scaling is very necessary in this case as we are dealing with distance based algorithm to extract the best performance from the model its better we do scaling before clustering. Also **PROBABILITY OF FULL PAYMENTS** is less than 0 and in single decimal places where as other columns are in units and tens ranges along with that the standard deviation of this column is very less compared to other columns since the variance is less between the columns Normalization technique is followed for scaling the data. Normalization follows (X-u)/s ,u is the mean and s is the standard deviation it converts the data to a scale of -1 to 1.To achieve this result Z-score from scipy.stats is used .When all the dimension are in the same range model works better in segregating the data points.

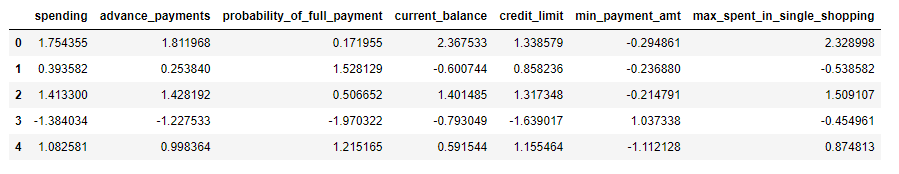


Table 5 Scaled Data

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

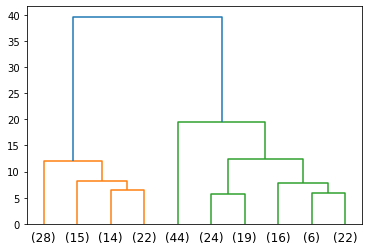


Figure 5 Dendogram

The above figure consists of dendogram , which is created using **ward** as the linkage method and affinity is **Euclidean. Ward’s** method works by the principle of how calculating the total distance from the centroids by adding two clusters , the clusters that shows minimum change are grouped together. **Euclidean** method calculates the distance between the clusters using the formula :

**√[(x2 – x1)2 + (y2 – y1)2]**.

The dendogram is the plot showing the distance between the clusters in y axis and each data points in x axis. The color coding in the dendogram shows 2 groups but looking at the business aspect we are trying to segregate the customers of a bank based on their past activities. So based on the business requirement we can choose the distance 15 that cuts the dendogram into 3 clusters so we can segregate the customers into high ,medium and low.

After finding the linkage the clusters are formed the labels are appended to the dataset and the following is the dataset with labels

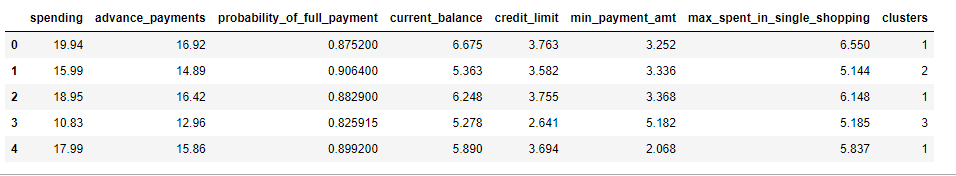


Table 6 Dataset with cluster labels

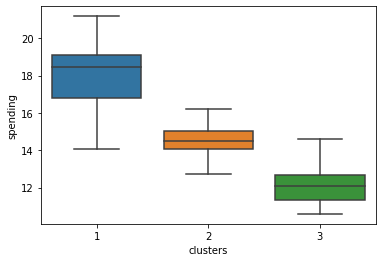


Figure 6. Boxplot of clusters based on spending

From the figure we can see that each cluster has different values of the median of spending , the 1st group consists of people spending more money ,the 2nd group spends in a medium scale and the 3rd one is the lowest.

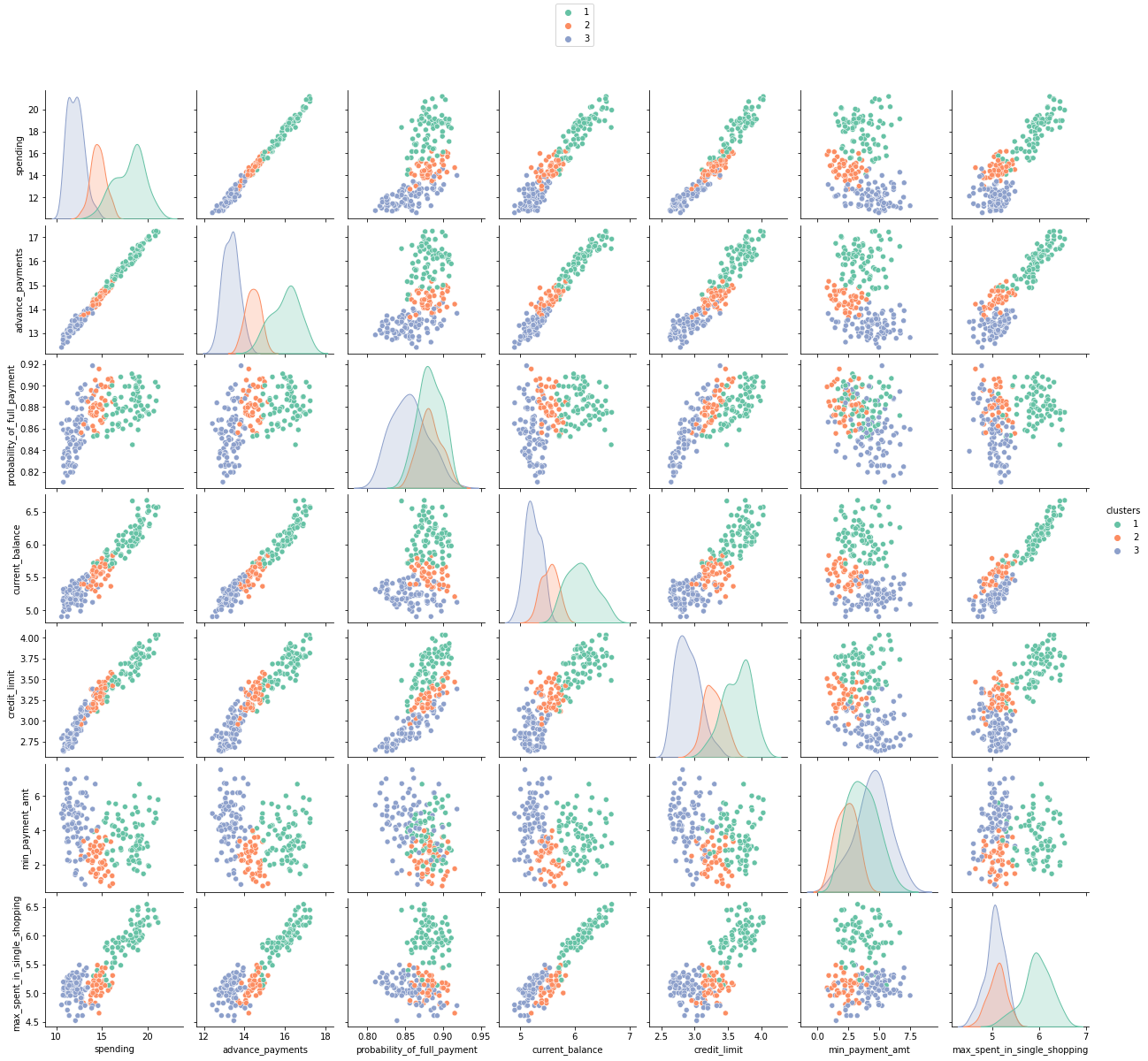


Figure 7 Pair plot of the clusters

Figure showing how each cluster is separated in each dimension.

## 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.

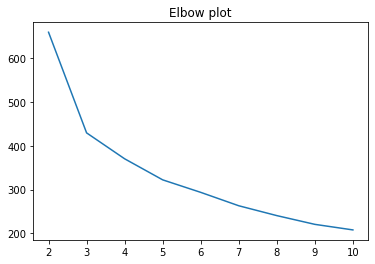


Figure 8 Elbow plot

The above figure consists of elbow plot, nothing but plotting the sum of squares within the clusters vs. the number of cluster. As we can see the value for wss starts reducing drastically from 2 to 3 and 3 to 4 after 4 the value flats out and reduces in a slower rate. The point at which stops reducing drastically is considered as the elbow point as it resembles elbow of the arm. This point indicates the correct number of clusters to be formed for this data set ,while looking at the plot we can say it flattens after 3 but the table below shows the exact values of wss for different clusters from which we can have a better conclusion that it flattens around 4.

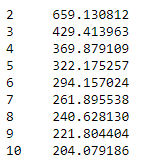


Table 6 Wss values

So hence we can conclude the optimum number of clusters for this dataset is 4 ,therefore customers can be divided into 4 groups.

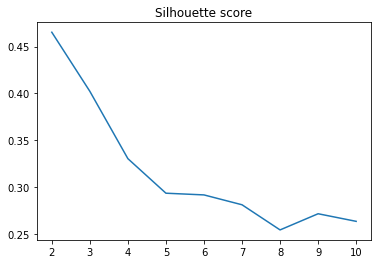


Figure 9 Silhouette score

The above graph shows the silhouette score for each cluster size we can see for cluster 4 the silhouette score 4 clusters is less than 3 and the exact silhouette score to be found out as **0.339337.**

The figure below shows the clusters separated with respect to each dimension as we can see from the pair plot the histograms have different peak among the cluster and the margin between the cluster is clear compared to the previous pair plot with cluster size equal to 3.

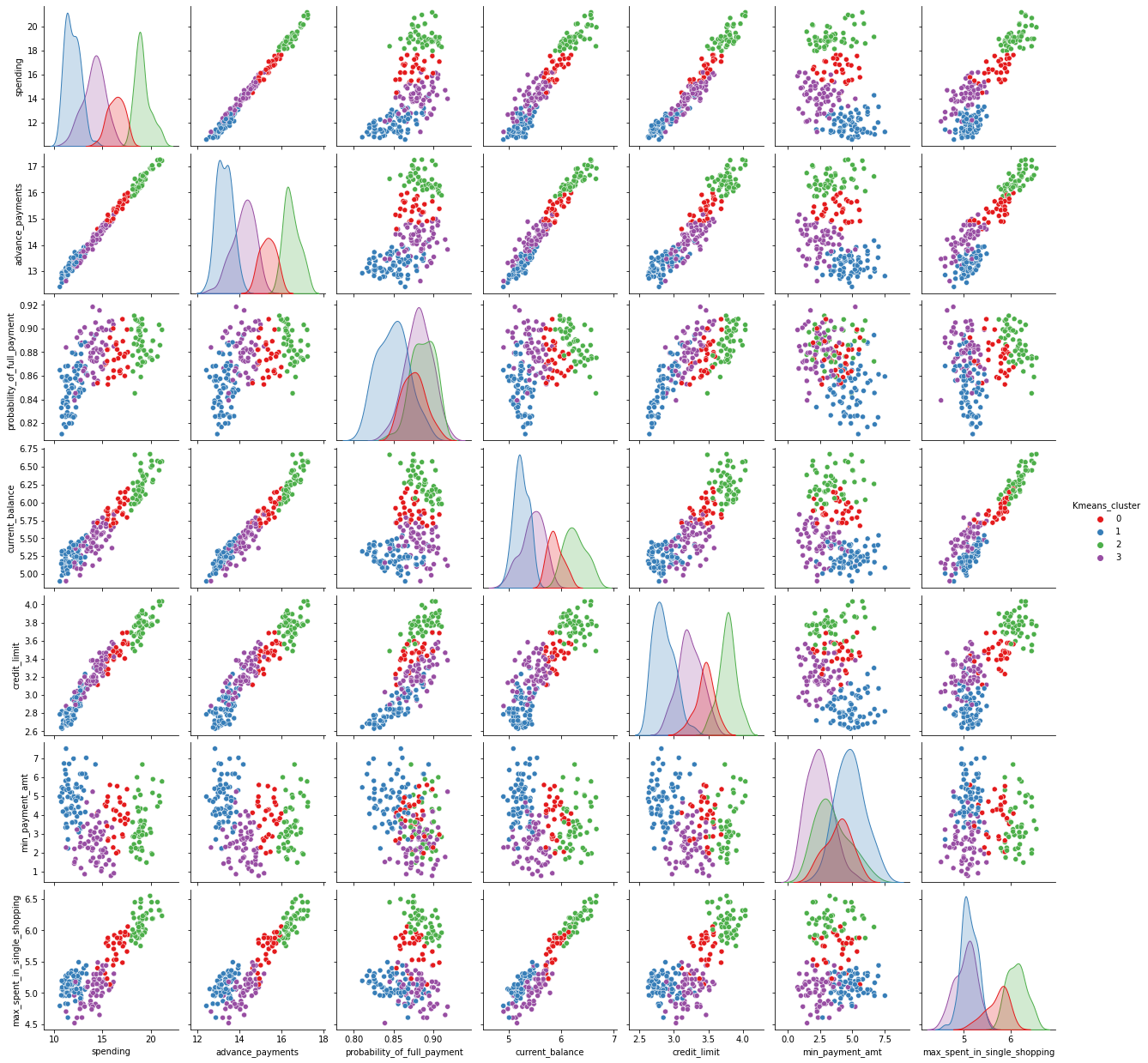
****

Figure. Pair plot of clusters

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

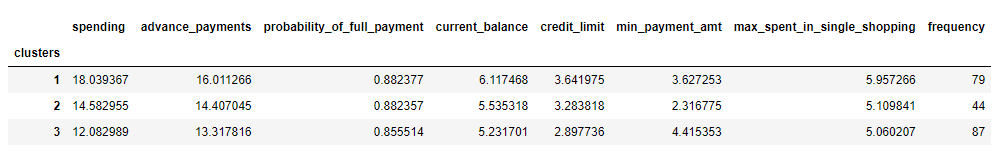


Table 7 Cluster Profiles

In average amount spent by the customers among the clusters differs ,for cluster 1 the average amount spent is the highest and for the 3rd cluster the amount spent is the least.

Advance payments also differs for each clusters , 3rd cluster with the minimum amount spent has the lowest advance payments ,2nd cluster has the medium advance payments and 1st cluster has the highest advance payments

Current balance is more for the 1st cluster with the most spending customers.

3rd cluster with minimum amount spending customers pay lot of minimum amount .

Based on the frequency there are more number of high spending or low spending customers

**BUSINESS RECOMMENDATIONS:**

We can reduce the initial payments for people who are spending less so that they will use credit card more to shop items which will increase the spending.

For people in 2nd cluster we can offer credit points if they spend the specified amount of money ,which they can use to buy items or return it to their account. This will drive the customers to spend more.

People who are spending more we can increase the credit limit so that it will allow them to spend more on shopping.

# 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.

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

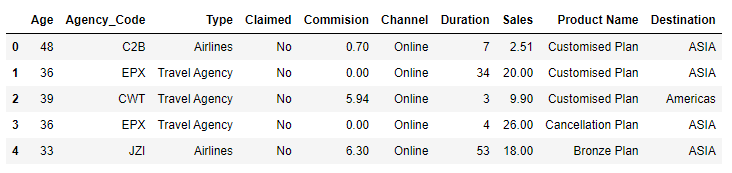


Table 8 Sample of the data

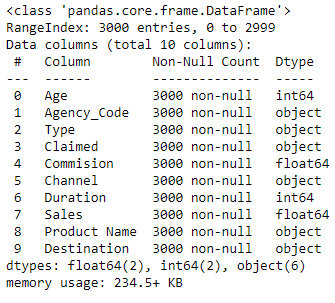


Table 9 Information of data

We can see from the above table there are 3000 rows and 10 columns ,there are 2 float type columns and 2 int type columns and 6 object type columns present in the dataset. In each column there are 3000 non null values which means there are no missing values present in the dataset.

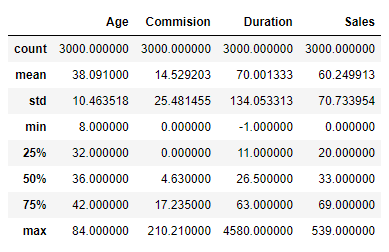


Table 10 Summary of the data

From the summary we can see there is -1 in duration ,logically speaking -1 can’t be the duration of a tour so we need to treat it before model creation. Along with that there are 139 duplicate values ,these values are removed.

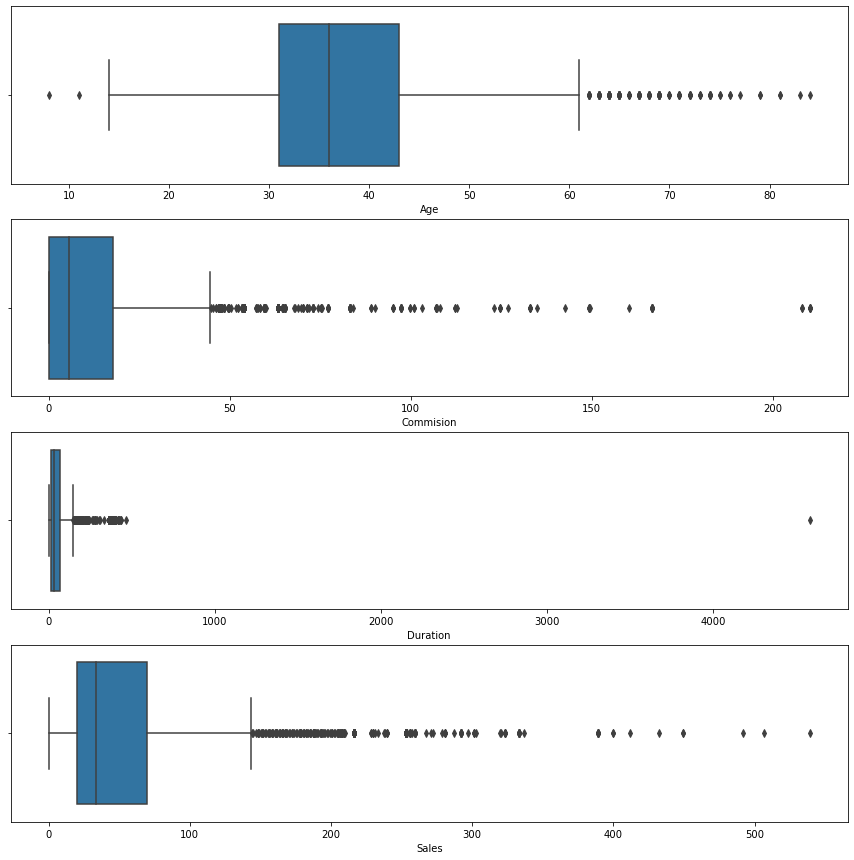


Figure 10 Boxplot of numerical columns

From the above figure we can see there are lot of outliers present in the data set, 41% of the data points are outliers.

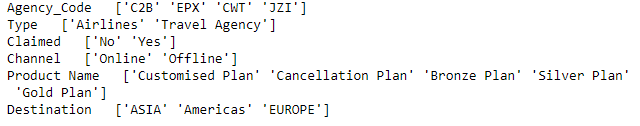


Figure 11 Unique values of categorical columns

From the figure above we can see there are no missing values or any incorrect values present in the dataset.

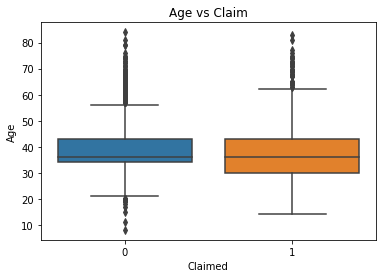


Figure 12 Boxplot on age

The plot shows the age doesn’t affect the claim rate both the class has the same median.

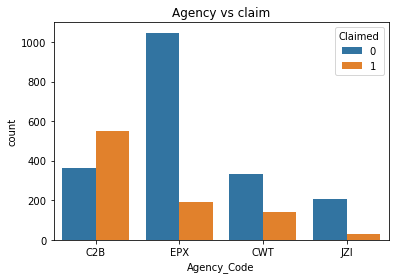


Figure 13 Agency code vs Claim

From the plot we can see that most of the claim is done by C2B agency and the least is by JZI.

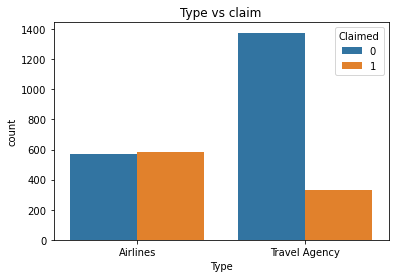


Figure 14 Type vs Claim

We can see most of the claim is by airlines and the least is by travel agency.

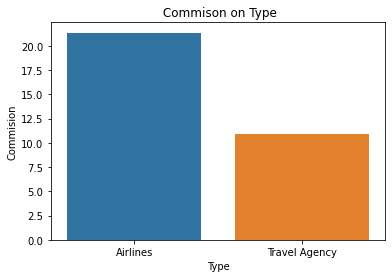


Figure 15 Commission vs. type

We can see the most of the commission is through airlines.

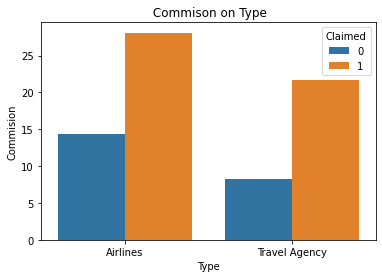


Figure 16 Commission by class

Most of the commission is from airlines through which claim is done.

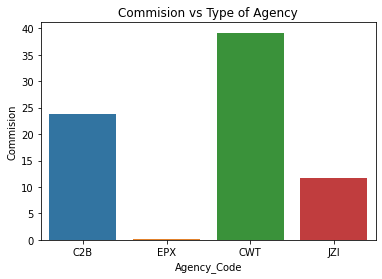


Figure 17 Commission vs. Agency

We can see that most of the commission is through CWT agency.

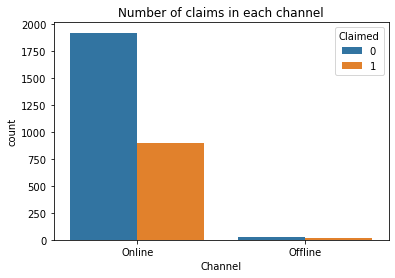


Figure 18 calims in each channel

Most of the claim is done through online channel.

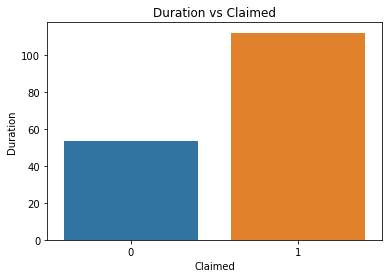


Figure 19 Duration vs. claim

Longer the duration of the travel higher the claim rate.

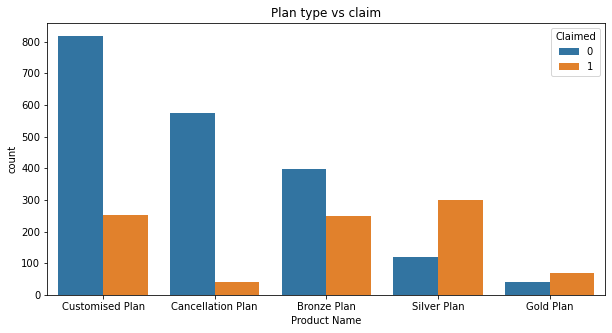


Figure 20 Plan type vs claim

Most of the claim is through Silver plan and Gold plan.

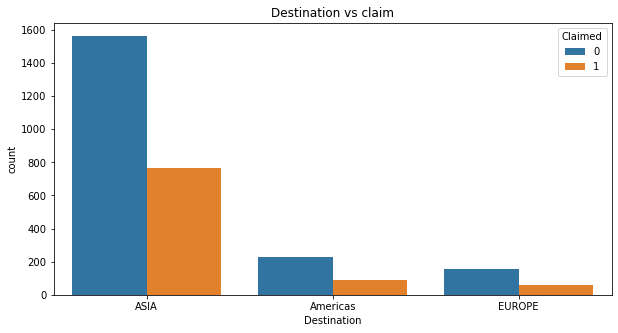


Figure 21 Destination vs. claim

Most of the claim are from travels which has destination as Asia ,Europe is the least of it.

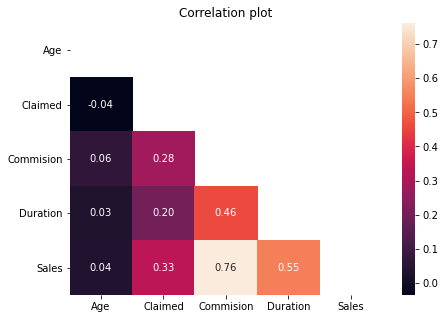


Figure 22

Sales has high positive correlation on Commission and Duration ,Age has negligible correlation.

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

After treating the incorrect values and converting the categorical columns into numerical columns ,data set is split using sklearn.train\_test\_split with the ratio of 70 30 percent the ratio of ones and zeros in train data set is given below:

0 0.678821

1 0.321179

and the ratio of 1s and 0s in the test data set is given below:

0 0.684517

1 0.315483

**Decision Tree:**

For the model the parameters chosen where max\_depth=6 , max\_feature= 4,random\_state=0.

Max\_depth : Decision tree is a greedy algorithm to avoid over fitting we must control the branch size , to control the branch size this parameter is used.

Max\_feature: to avoid over fitting and to make the model optimized only 4 feature is allowed to be chosen to calculate the gini at every node.

Random\_state: to reproduce the same result every time we run the model

**Random Forest:**

Parameters chosen where n\_estimators=500,max\_depth=8,max\_feature=4

n\_estimator: total number of decision tree in the model

**Neural Network:**

Parameters activation =relu,hidden layers=500,500,500,max\_iter=1000,solver=adam,tot=0.1

Activation: function used to activate each neurons in the model

Hidden layers: structure of the neural network

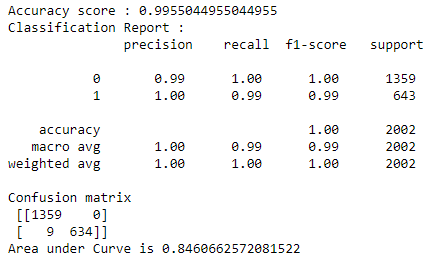
Max\_iter: maximum number of iterations done to train the model

Solver: algorithm to calculate the error and find the right variables in the parameter space

Tot: learning rate, rate at which the neural net trains

## 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.

**Decision Tree: Train dataset**



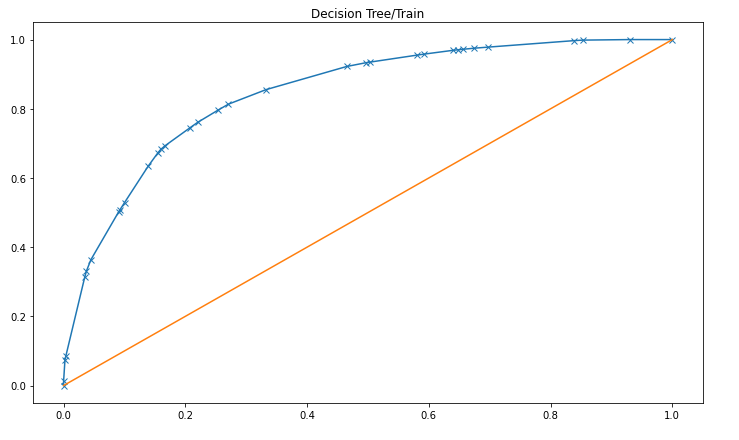
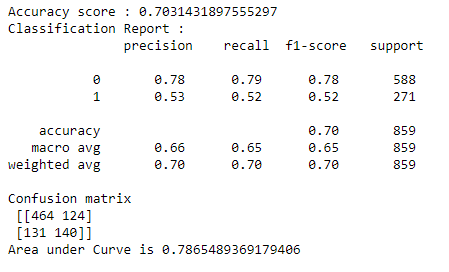


Figure 23 Roc curve for decison tree

**Test dataset:**



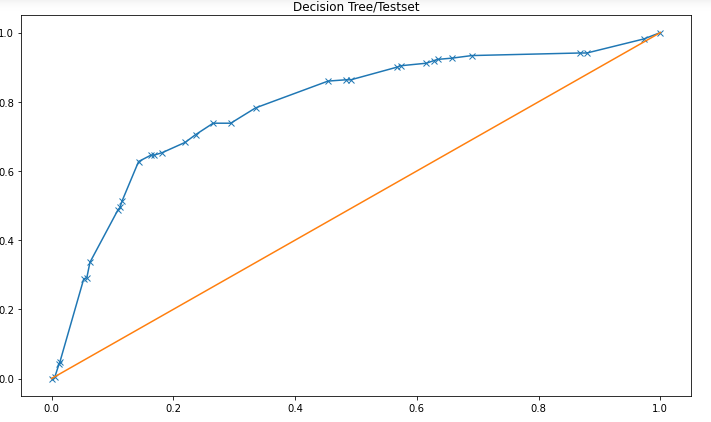
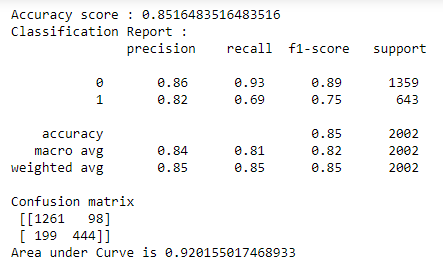


Figure 24 Roc curve test dataset

Model is over fitted performs better on training set but not in the test data.f1 score is also not really good accuracy is around 70 % area under the curve is also 0.78

**Random Forest: Train dataset**



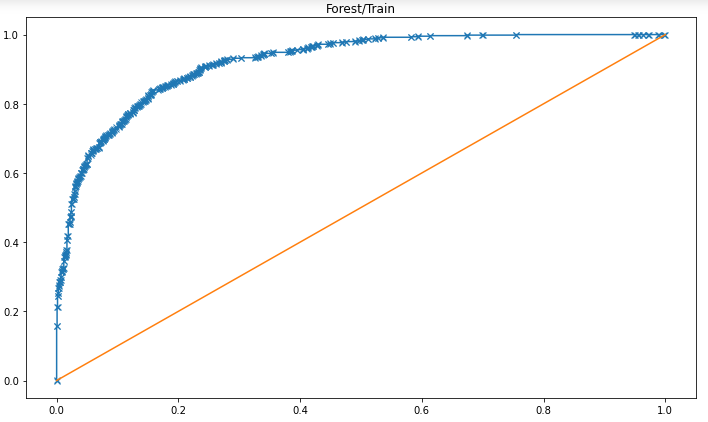
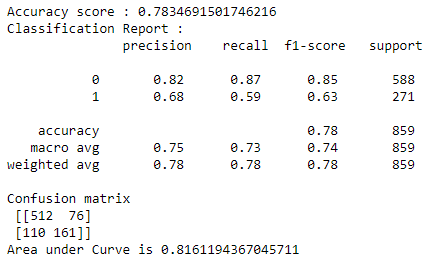


Figure 25 Roc curve for train dataset

**Test dataset:**



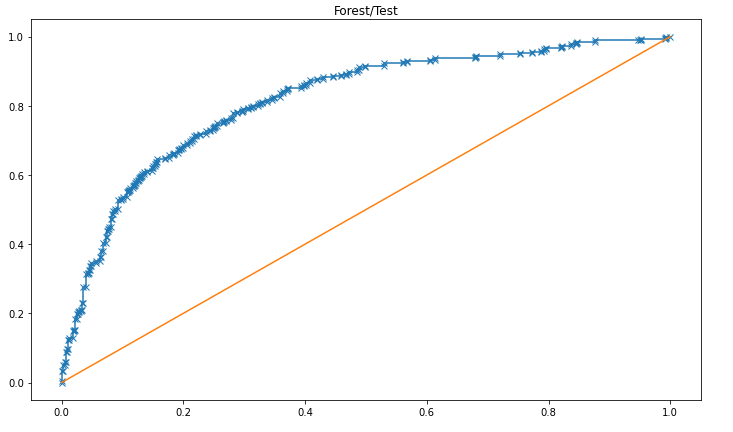
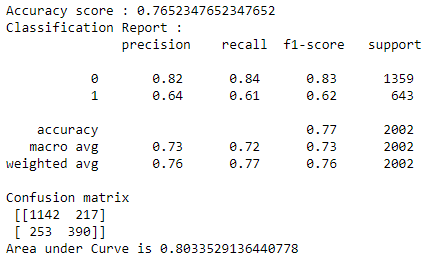


Figure 26 Roc curve for test set

Model has a good true positive rate and true negative rate, model preforms better in both training and test dataset. Area under the curve is better than decision tree.F1 score is also better than decision tree.

**Neural Network: Train**



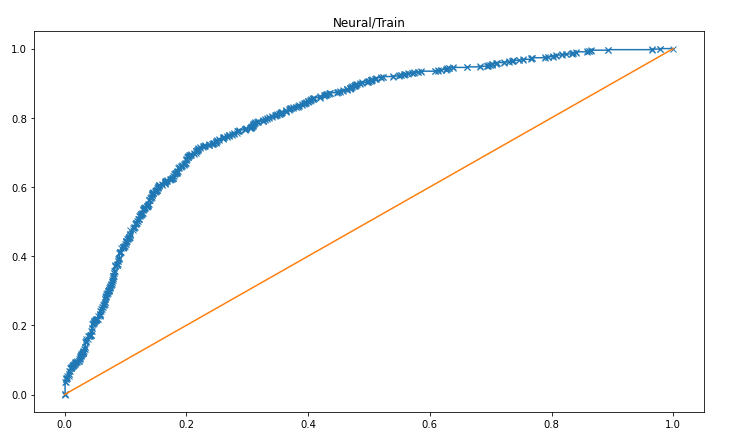
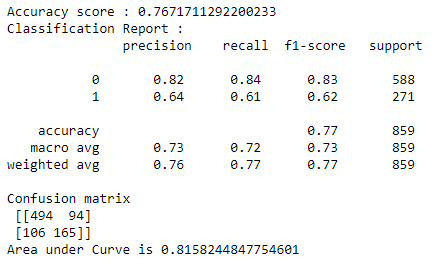


Figure 27 Roc curve Ann Train data

**Test Dataset:**



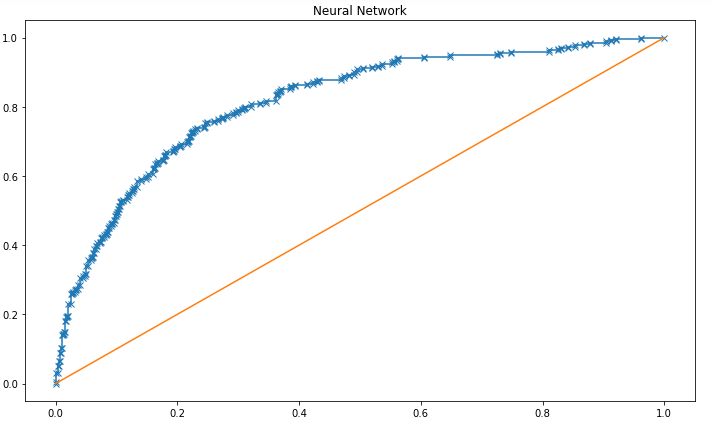


Figure 28 Roc curve ann test data

Model works better in test data than training data it has generalized properly and f1 score, auc is similar to random forest overall the model is efficient.

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

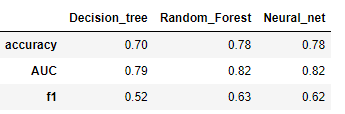
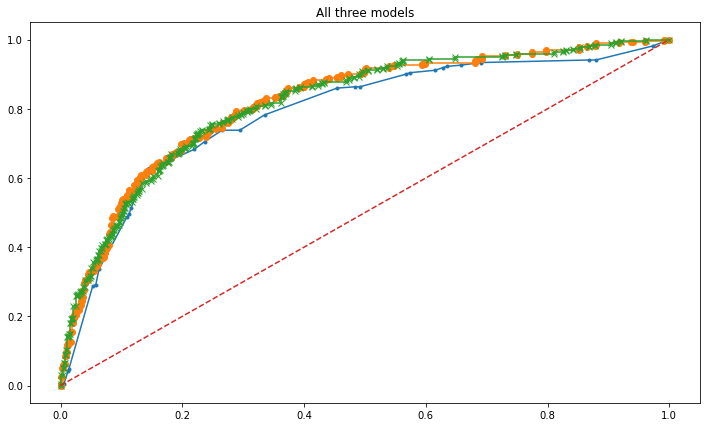


Table 1 Performance metrics

From the above table we can see that decision tree has the least performance in all measures compared to other two models. Random forest and neural nets works more or less equally random forest has higher f1 score than neural network so random forest is the best model in this case.



X-neural network

0 –random forest

.-decision tree

Figure 2 Roc curve for all 3 models

From the plot we can see the roc curve is also similar for random forest and neural network.

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

**Insights:**

* Age doesn’t affect the claim
* Most of the claim is done byC2B agency and least by JZI
* Most of the claim is by airways travel
* We have the most of the claim from the high commission travels by airlines
* Most of the claims are in silver plan and gold plan
* Most of the claims are from Asia destination

**Recommendations:**

* Changing the plans in gold and silver plans might help reducing the claims
* Should have a survey on the C2B agency which might en-light us why most of the claim is from that agency and we can rectify if there is any foul practices.
* Changing the policies of the insurance for a particular destination like asia might reduce the claims