

AUTOMOBILE

DATA

ANALYSIS

-

Siddhant Raj

Singh

Date: 21St Sep, 2022

Table of Contents

Contents

Problem 1

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

1.2 Do you think scaling is necessary for clustering in this case? Justify)………………………………………………………………………………………………………………………………………12

1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them…………………………………………………………………………………….13

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. )……………………………………………………………………………………………………………………………………20

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.…………………………………………………………………………………………………………………24

Problem 2

2.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariateanalysis)........................................................................................................................... 29

2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network ............................................................................................................................................................ 40

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 .......................................................................................................................................... 48

2.4 Final Model: Compare all the models and write an inference which model is best/optimized.................................................................................................................................... 62

2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations .............................................................................................................................................................64

List of Figures

|  |
| --- |
| Class\_Report\_Train\_RF…………………………………………………………………………………………………………………………………57  x\_rfcl.png …………………………………………………………………………………………………………………………………48  agg\_mean\_clust\_w\_freq.png……………………………………………………………………………………………………………………………26  agg\_mean\_clust\_w.png……………………………………………………………………………………………………………………………27  agg\_mean\_clust.png…………………………………………………………………………………………………………………………………18  bank\_marketing\_part1\_Data.csv  Bi\_Heat\_map.png……………………………………………………………………………………………………………………………13  Bi\_Pair\_plot.png…………………………………………………………………………………………………………………………………11  boxplot\_Agency\_code\_claimed.png………………………………………………………………………………………………………………36  boxplot\_Channel\_Claimed.png………………………………………………………………………………………………………………………39  boxplot\_Destination\_Claimed.png…………………………………………………………………………………………………………………40  boxplot\_ins.png…………………………………………………………………………………………………………………………………34  boxplot\_Product\_claimed.png………………………………………………………………………………………………………………………37  boxplot\_Type.png…………………………………………………………………………………………………………………………………38  Class\_Report\_Test\_DT…………………………………………………………………………………………………………………………………52  Class\_Report\_Test\_nncl …………………………………………………………………………………………………………………………………62  Class\_Report\_Test\_Rf …………………………………………………………………………………………………………………………………57  Class\_Report\_Train\_DT …………………………………………………………………………………………………………………………………52  Class\_Report\_Train\_nncl …………………………………………………………………………………………………………………………………61  clust\_inertia.png………………………………………………………………………………………………………………………………23  clust\_profile\_k.png…………………………………………………………………………………………………………………………………28  clusters\_boxplot\_w.png……………………………………………………………………………………………………………………………27  confusion\_matrix\_test\_DT.png………………………………………………………………………………………………………………………51  confusion\_matrix\_test\_nncl.png………………………………………………………………………………………………………………………60  confusion\_matrix\_test\_Rf.png………………………………………………………………………………………………………………………56  confusion\_matrix\_train\_DT.png…………………………………………………………………………………………………………………50  confusion\_matrix\_train\_nncl.png……………………………………………………………………………………………………………………60  confusion\_matrix\_train\_Rf.png……………………………………………………………………………………………………………………55  countplot\_Agency\_code.png………………………………………………………………………………………………………………………36  countplot\_Channel.png…………………………………………………………………………………………………………………………………39  countplot\_Destination.png…………………………………………………………………………………………………………………………………40  countplot\_Product.png…………………………………………………………………………………………………………………………………37  countplot\_Sales.png…………………………………………………………………………………………………………………………………36  countplot\_Type.png…………………………………………………………………………………………………………………………………38  D\_tree.pdf  D\_tree.png  data.png  Dendrogram\_avg.png…………………………………………………………………………………………………………………………………16  dendrogram\_trunc\_10\_w.png…………………………………………………………………………………………………………………………19  dendrogram\_trunc\_10.png……………………………………………………………………………………………………………………………17  dendrogram\_trunc\_20\_w.png………………………………………………………………………………………………………………………19  dendrogram\_trunc\_20.png……………………………………………………………………………………………………………………………17  Dendrogram\_wards.png…………………………………………………………………………………………………………………………………18  df\_desc.png…………………………………………………………………………………………………………………………………6  df\_report\_test…………………………………………………………………………………………………………………………………56  df\_report\_test\_nncl.png…………………………………………………………………………………………………………………………………62  df\_report\_test\_rf.png…………………………………………………………………………………………………………………………………57  df\_report\_test.png…………………………………………………………………………………………………………………………………56  df\_report\_train\_nncl.png…………………………………………………………………………………………………………………………………61  df\_report\_train\_rf.png…………………………………………………………………………………………………………………………………611  df\_report\_train.png…………………………………………………………………………………………………………………………………56  displot\_ins.png…………………………………………………………………………………………………………………………………35  distplot\_sales.png…………………………………………………………………………………………………………………………………69  Feat\_imp\_bar\_rfcl.png…………………………………………………………………………………………………………………………………48  Feat\_imp\_bar.png…………………………………………………………………………………………………………………………………46  heatmap.png…………………………………………………………………………………………………………………………………42  ins\_df\_desc.png…………………………………………………………………………………………………………………………………32  ins\_df.png  ins\_pairplot.png…………………………………………………………………………………………………………………………………41  ROC\_3models\_test.png…………………………………………………………………………………………………………………………………65  ROC\_3models\_train.png…………………………………………………………………………………………………………………………………65  ROC\_test\_DT.png…………………………………………………………………………………………………………………………………53  ROC\_test\_nncl.png…………………………………………………………………………………………………………………………………63  ROC\_test\_Rf.png…………………………………………………………………………………………………………………………………58  ROC\_train\_DT.png…………………………………………………………………………………………………………………………………53  ROC\_train\_nncl.png…………………………………………………………………………………………………………………………………63  ROC\_train\_Rf.png…………………………………………………………………………………………………………………………………58  sample\_data.png…………………………………………………………………………………………………………………………………5  sample\_ins.png…………………………………………………………………………………………………………………………………31  scatter\_clust\_k\_means.png……………………………………………………………………………………………………………………………29  scatter\_clust.png…………………………………………………………………………………………………………………………………27  tree\_regularised.dot  Uni\_Boxplot.png…………………………………………………………………………………………………………………………………8  Uni\_Hist.png…………………………………………………………………………………………………………………………………9  wss\_df.png…………………………………………………………………………………………………………………………………23 |

# **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. Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month. 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).

# 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:

Table

Description automatically generated

Table 1. Dataset Sample

* Shape of the data is (210, 7) i.e. 7 columns and 210 rows.
* The sample of the dataset i.e. the first five rows appear to be perfect.
* After checking and gathering more information about the data, we find that there are total 210 entries in the dataset and all columns are of data type float64.
* The dataset has no null values.
* The dataset has no duplicate values.

# Exploratory Data Analysis

## Let us check the types of variables in the data frame.

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 spending 210 non-null float64

1 advance\_payments 210 non-null float64

2 probability\_of\_full\_payment 210 non-null float64

3 current\_balance 210 non-null float64

4 credit\_limit 210 non-null float64

5 min\_payment\_amt 210 non-null float64

6 max\_spent\_in\_single\_shopping 210 non-null float64

There are total 210 rows and 7 columns in the dataset. All the columns are of float64 type.

## Check for missing values in the dataset:

From the above results we can see that there are no missing values present in the dataset. There are total 210 rows in the data set and all the columns have 210 non-null values.

## Describing the Data:

Table

Description automatically generated

* The variable ‘Spending’ tells the amount spent by the customer per month (in 1000s). From the above table we can conclude that the average amount spent by the customer per month is 14.847524 with a standard deviation of 2.909699. The median for the same is 14.35500 which is slightly less than the mean. Therefore we can say that the variable ‘spending’ is slightly skewed to the right. The maximum and minimum amount spent by the customer is 21.1800 and 10.5900 respectively.
* The variable ‘advance\_payments’ tells the amount paid by the customer in advance by cash (in 100s). From the above table we can conclude that the average amount paid by the customer in advance by cash is 14.559286 with a standard deviation of 1.305959. The median for the same is 14.32 which is slightly less than the mean. Therefore we can say that the variable ‘advance\_payments’ is slightly skewed to the right. The maximum and minimum amount paid by the customer in advance by cash is 17.2500 and 12.4100 respectively.
* The variable ‘probability\_of\_full\_payment’ tells the probability of payment done in full by the customer to the bank. From the above table we can conclude that the average probability of payment done in full by the customer is 0.870999 with a standard deviation of 0.023629. The median for the same is 0.87345 which is slightly more than the mean. Therefore we can say that the variable ‘probability\_of\_full\_payment’ is slightly skewed to the left. The maximum and minimum probability is 0.9183 and 0.8081 respectively.
* The variable ‘current\_balance’ tells the balance amount left in the account to make purchases (in 1000s). From the above table we can conclude that the average amount left in the account is 5.628533 with a standard deviation of 0.443063. The median for the same is 5.52350 which is slightly less than the mean. Therefore we can say that the variable ‘current\_balance’ is slightly skewed to the right.
* The variable ‘credit\_limit’ tells limit of the amount in the credit card (in 10000s). From the above table we can conclude that the average limit of the amount in the credit card is 3.258605. The median for the same is 3.23700n which is slightly less than the mean. Therefore we can say that the variable ‘credit\_limit’ is slightly skewed to the right.
* The variable ‘min\_payment\_amt’ tells minimum amount paid by the customer while making payments for purchases made monthly (in 100s). The mean is 3.700201 and median is 3.59900. Since mean is greater than the median we can say that the variable ‘min\_payment\_amt’ is slightly skewed to the right. The maximum and minimum amount paid by the customer for the same is 8.4560 and 0.7651 respectively. This also suggests that data is widely spread.

## Univariate Analysis

## Box Plot

Chart, box and whisker chart

Description automatically generated

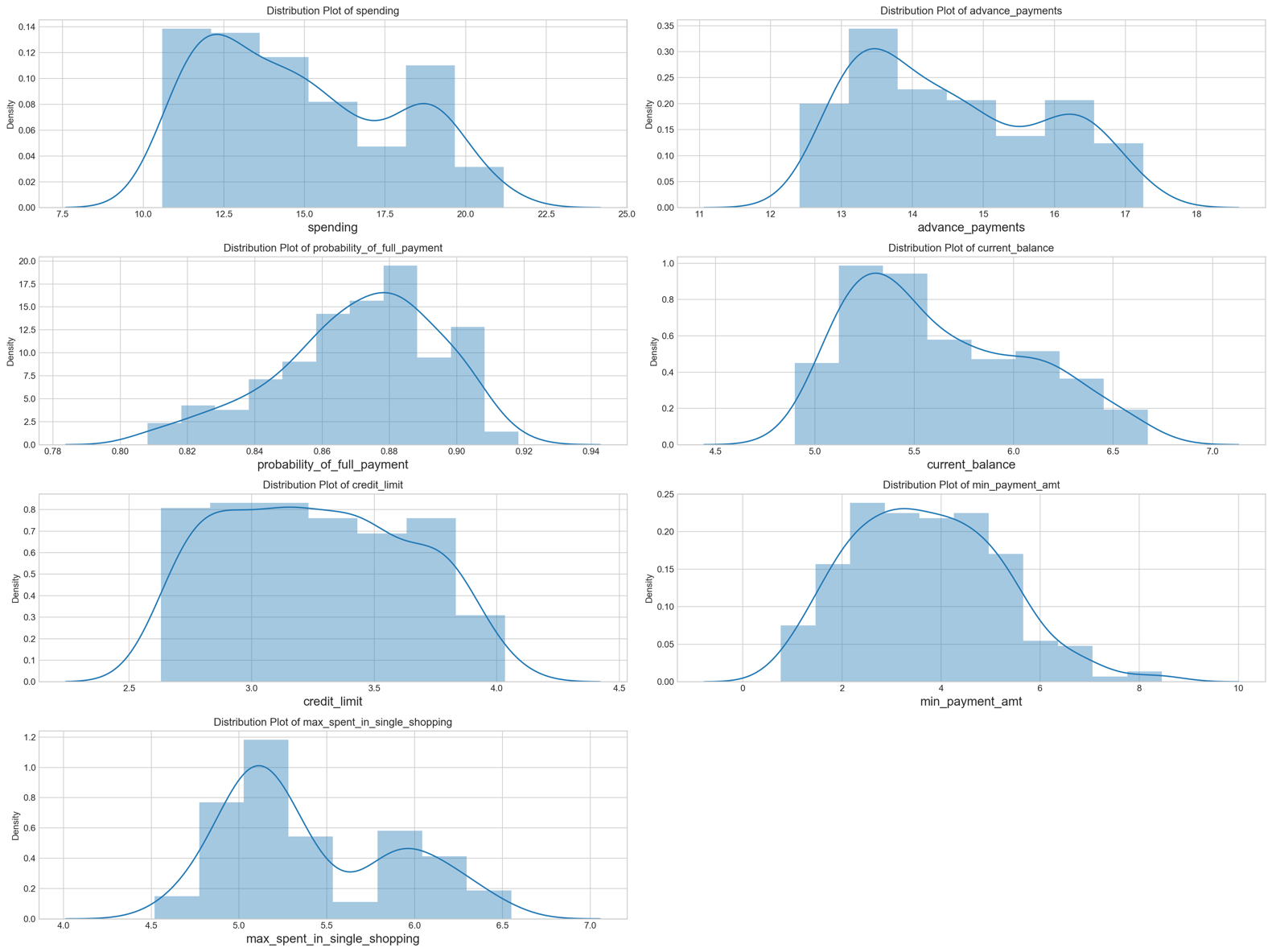
* The variable ‘spending’ is slightly skewed to the right. The median number of ‘spending’ is approximately between 14 and 15. The Inter-Quartile range is about 5 which means about 50% of ‘spending’ is approximately between 12.5 and 17.5. The variable ‘spending’ has no outliers.
* The variable ‘advance\_payments’ is slightly skewed to the right. The median number of ‘advance\_payments’ is approximately between 14 and 14.5. The Inter-Quartile range is about 2.25 which means about 50% of ‘advance\_payments’ is approximately between 13.5 and 15.75. The variable ‘advance payments’ has no outliers.
* The variable ‘probability\_of\_full\_payment’ is slightly skewed to the left. The median is approximately 0.87. The Inter-Quartile range is about 0.30 which means about 50% of ‘probability\_of\_full\_payment’ is approximately between 0.855 and 0.885. The variable has few outliers.
* The variable ‘min\_payment\_amt’ is slightly skewed to the right. From the above figure we can conclude that there are outliers for the variable ‘min\_payment\_amt’ which tells the minimum amount paid by the customer while making payments for purchases made monthly. This suggests that there are few customers who had to pay higher minimum amount while making payments because their purchases were relatively expensive as compared to other customers.

The median is approximately 3.60. The Inter-Quartile range is about 2.20 which means about 50% of ‘min\_payment\_amt’ is approximately between 2.55 and 4.75.

* The variable ‘credit\_limit’ is slightly skewed to the right. Its almost Normally distributed. The median number of ‘credit\_limit’ is approximately 3.25. The variable ‘credit\_limit’ has no outliers. The inter quartile range is about 0.60 which means about 50% of ‘credit\_limit’ is approximately between 2.95 and 3.55.
* The variable ‘current\_balance’ is slightly skewed to the right. It is almost normally distributed. The median number of ‘current\_balance’ is approximately 5.55. The variable ‘current\_balance’ has no outlier. The inter quartile range is about 0.70 which means about 50% of ‘current\_balance’ is approximately between 5.26 and 5.96.
* By looking at the visual distribution of all the variables using a Box Plot, we can say that most of the information derived about the variables are suggestive from ‘Describing the Data’ section.

## 

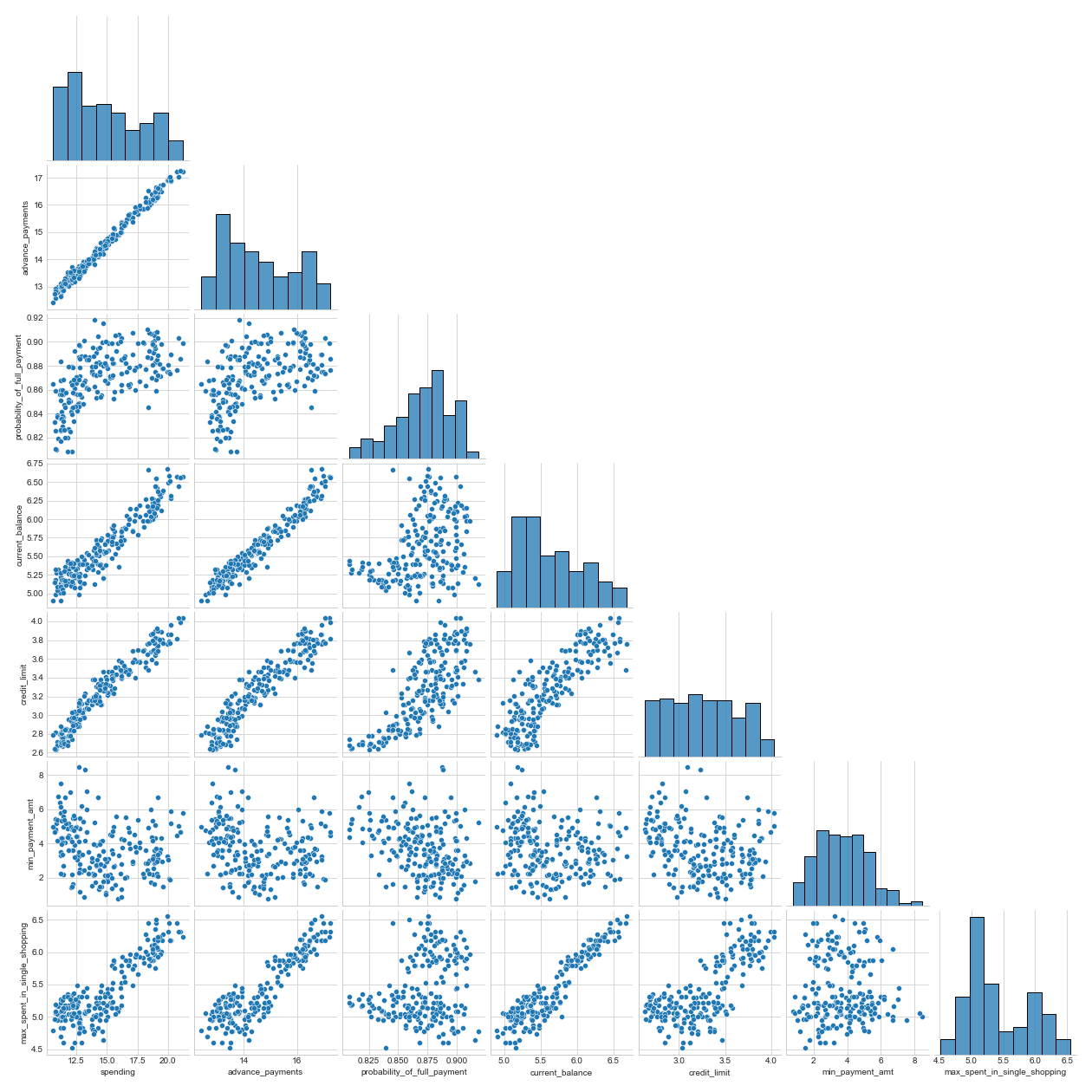
## Distribution Plot



* For the variable ‘Spending’ we can see that there could be chance of multi modes in the dataset. The dist. plot shows the distribution of data from 10 to 22. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘advance\_payments’ we can see that there could be a chance of multi nodes In the dataset. The dist. plot shows the distribution of data from 12 to 17. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘probability\_of\_full\_payment’ we can see that the distribution plot shows the distribution of data from 0.80 to 0.92. We can clearly see from the distribution plot that the variable is negatively skewed.
* For the variable ‘current\_balance’ we can see that the distribution plot shows the data from 5.0 to 6.5. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘credit\_limit’ we can see that the distribution plot shows the data from 2.5 to 4.0. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘min\_payment\_amt’ we can see that the distribution plot shows the data from 2 to 8. We can clearly see from the distribution plot that the variable is positively skewed.
* For the variable ‘max\_spent\_in\_single\_shopping’ we can see that the distribution plot shows the data from 4.5 to 6.5. We can clearly see from the distribution plot that the variable is positively skewed.

## Bivariate Analysis

## Pair Plot



* The above Pairplot shows the relationship between the variables present in the Data Frame using scatter plot and distribution of the variable using Histogram.
* There is a strong positive correlation between the variable ‘spending’ and ‘advance\_payments’. This means that higher the amount spent by the customer per month, more the customer is able to pay in advance by cash.
* There is a strong positive correlation between the variable ‘current\_balance’ and ‘max\_spent\_in\_single\_shopping’. This means that higher the balance amount left in the account to make purchases, higher the amount spent in one purchase. The customer plans to spend the amount based on the current balance amount available.
* There is a strong positive correlation between the variable ‘current\_balance’ and ‘advance\_payments’. This means that higher the balance amount left in the account to make purchases, higher the amount paid by the customer in advance in cash.
* There is a strong positive correlation between the variable ‘credit\_limit’ and ‘spending’. This means that higher the limit of amount in credit card, more is the amount spent by the customer per month.
* There is a positive correlation between the variable ‘probability\_of\_full\_payment ’ and ‘credit limit’. This means that higher the limit of amount in credit card, higher the probability of payment done in full by the customer to the bank.
* There is a positive correlation between the variable ‘current\_balance’ and ‘spending’. This means that higher the balance amount left in the account to make purchases, higher the amount spent by the customer per month.
* There is some amount of positive correlation between the variable ‘probability\_of\_full\_payment’ and ‘current\_balance’. This means that as we increase the ‘current\_balance’, the ‘probability\_of\_full\_payment’ also increases but after a certain point the ‘current\_balance’ increases without showing much increase in the ‘probability\_of\_full\_payment’.
* There is a strong positive correlation between the variable ‘current\_balance’ and ‘credit\_limit’. This means that higher the balance amount left in the account to make purchases, higher the limit of amount in credit card.
* There is a strong positive correlation between the variable ‘advance\_payments’ and ‘credit\_limit’. This means that higher the limit in the account to make purchases, higher the amounts paid by the customer in advance by cash.
* The variable ‘min\_payment\_amt’ is not correlated to any of the variables, and therefore it is not affected by any changes in the ‘current\_balance’ or ‘credit\_limit’ of the customers.

## Heat Map

Chart

Description automatically generated with medium confidence

# 

From the above Heat Map we can observe the following points:

* ‘spending’ and ‘advance\_payments’ are strongly correlated (0.99).
* ‘max\_spent\_in\_single\_shopping’ and ‘current\_balance’ are strongly correlated (0.93).
* ‘current\_balance’ and ‘advance\_payments’ are strongly correlated (0.97).
* ‘current\_balance’ and ‘spending’ are strongly correlated (0.95).
* ‘credit\_limit’ and ‘spending’ are strongly correlated (0.97).
* ‘probability\_of\_full\_payment’ and ‘credit\_limit’ are positively correlated (0.76)
* ‘probability\_of\_full\_payment’ and ‘current\_balance’ are somewhat positively correlated (0.37). The ‘probability\_of\_full\_payment’ stops increasing when we reach a certain point in the ‘current\_balance’.
* ‘current\_balance’ and ‘credit\_limit’ are positively correlated (0.86).
* ‘advance\_payments’ and ‘credit\_limit’ is strongly positively correlated (0.94).
* ‘min\_payment\_amt’ is negatively correlated with to all of the other variables given in the dataset.

From the above two graphs i.e. Pair plot and Heat Map we can conclude that the customers having higher current balance amount in their account are the ones making highest spendings as well as advance payments in cash and therefore they have higher limits of amount in credit card (since they are able to make good financial decisions). Since these customers have high credit limit, majority of them are having highest probability of making the full payment.

**NOTE :** We already know that if there are outliers present in our datasets then they should be treated. This is because clustering results are affected by the presence of outliers. Therefore we have treated the outlier by creating a function called as ‘remove\_outliers( )’. The function takes in a column as an argument and calculates the IQR (Inter-Quartile Range) of that column. Using the Inter-Quartile Range we calculate the Upper-Limit and the Lower-Limit of the column and remove the outliers using the same.

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

Yes, I think that Scaling is necessary and is one of the most important steps while performing Clustering.

This is because Clustering tries to identify groups which are similar within objects but are highly dissimilar between other objects i.e. homogeneity within the groups heterogeneity between the groups. The Clustering model works on the distance based computations. If we don’t normalize the features, we will end up giving more weight to some features than the others. Differences in the scales across input variables may increase the difficulty of the problem being modelled. This is the reason why scaling needs to be performed.

In the dataset provided we have following features : spending, advance payments, probability of full payment, current balance, credit limit, minimum payment amount and maximum amount spent in a single shopping.

Following is maximum, minimum, standard deviation and variance of the feature ‘spending’:

(21.18, 10.59, 2.902763307757227, 8.426034820861677)

Following is maximum, minimum, standard deviation and variance of the feature ‘advance\_payments’:

(17.25, 12.41, 1.30284559048763, 1.6974066326530612)

Following is maximum, minimum, standard deviation and variance of the feature ‘probability\_of\_full\_payment’:

(0.9183, 0.8105875, 0.023503842501762292, 0.0005524306123476474)

We can see that the maximum, minimum, standard deviation and variance of ‘spending’ or ’advance\_payments’ is different as compared to that of ‘probability\_of\_full\_payment’. The standard deviation of ‘spending’ is 2.9 and for ‘probability\_of\_full\_payment’ it is 0.02. If we don’t scale the feature ‘probability\_of\_full\_payment’ then it will be outweighed by other features like ‘spending’ or ‘advance\_payments’ and hence It is important to scale the data.

I have used standard scaler for scaling the data. Using standard scaler method we standardize the dataset. It involves rescaling the distribution of values so that the mean of observed values is 0 and the standard deviation is 1.

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

**Hierarchical Clustering**

* Hierarchical Clustering is the clustering technique in which records are grouped to create clusters based on distances between records and distances between clusters. We try to achieve heterogeneity between the groups and homogeneity within the groups. There are two types of approach that we take in hierarchical clustering.
  + Agglomerative approach
  + Divisive approach.
* The similar records can be grouped using the following Distances measure:
  + Euclidian Distance
  + Manhattan Distance
  + Chebyshev Distance
  + Minkowski Distance
* The similar clusters can be grouped using the following Linkage types:
  + Single Linkage
  + Complete Linkage
  + Average Linkage
  + Centroid Linkage
  + Wards Linkage

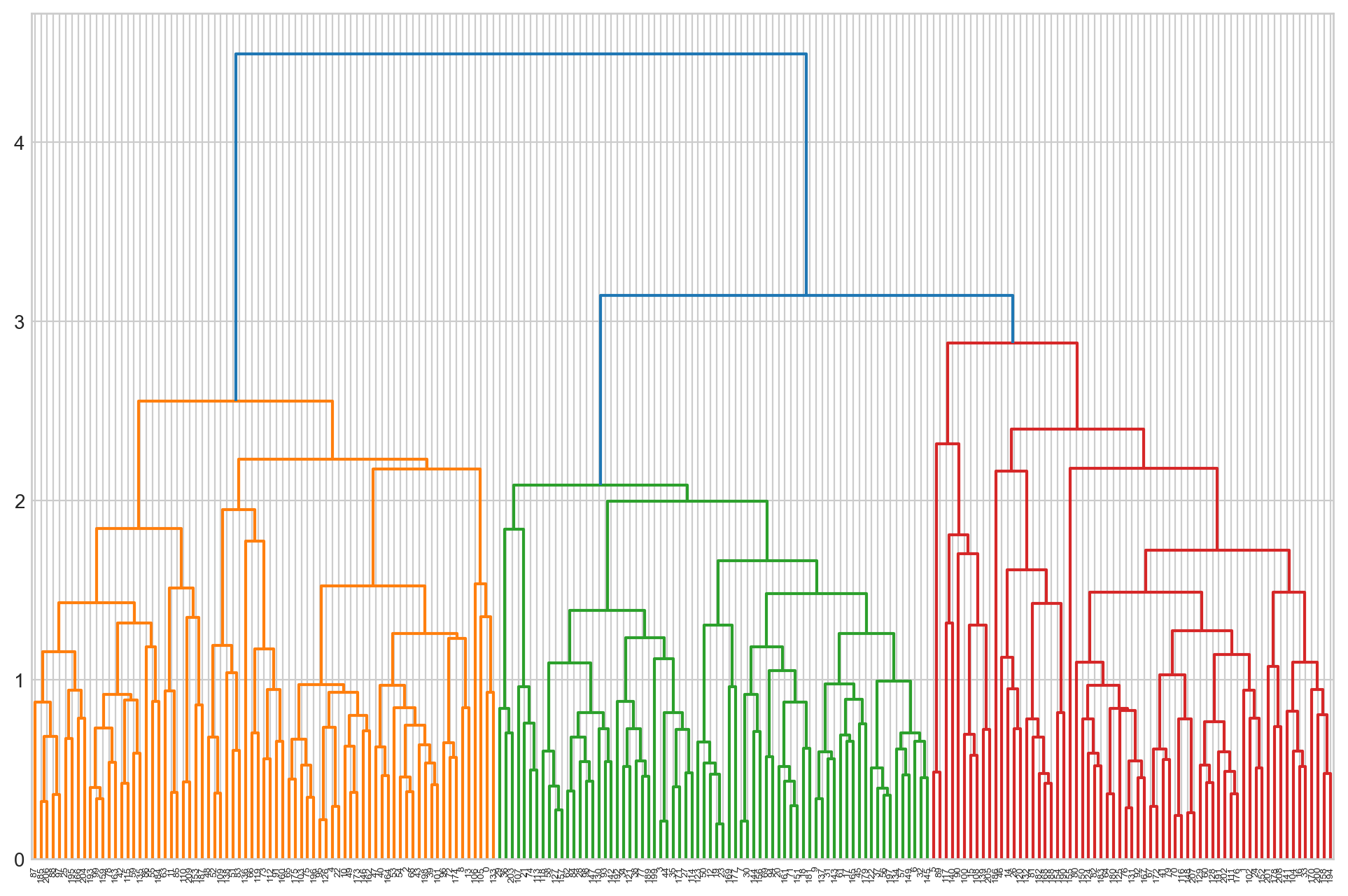
**Dendrogram**

* A dendrogram is tree like diagram that summarizes the process of clustering.
* On the x axis are the records of the data.
* Similar records are joined using vertical lines whose vertical length reflects the distances between the records. Greater the difference in height, more the dissimilarity.
* By choosing a cut-off distance on the y-axis, a set of clusters is chosen.

# 

**NOTE**: We have used Euclidian Distance as the Distance measure and   
Average Linkage and Wards Linkage as Linkage type.

**Average Linkage Method**



* We imported from the scientific python (submodule hierarchy), the two functions called as linkage and dendrogram.
* Dendrogram is for visualisation, linkage is for computing the distances and merging the clusters from n to 1.
* First we created linkage using the Average Linkage method. This linkage method will store the various distances at which the n cluster are sequentially merged into a single cluster.
* Once we store the linkage information, the next step is to visualise the data. For this we will call the dendrogram function.
* In the beginning, the dendrogram that is created is not truncated. We observe that dendrogram creates a partition for us and gives us a colour code for the same. Total number of clusters that the dendrogram is suggesting to us is 3 with colour code as orange, green and red.

Chart

Description automatically generated

* We then pass additional parameters which truncates the dendrogram. We pass a parameter called as 'truncate\_mode' which is set equal to 'lastp'. This will ensure that only the last 'p' clusters that are merged is shown to us within the visual. We first set the 'p' value as 20. We get a dendrogram which shows the last 20 merges. Every horizontal line is a merge. We repeat the step with 'p' value set to 10.

Chart

Description automatically generated

* On observing the dendrogram, we identify a cut-off point. We choose the cut-off where the jump or vertical lines should be of highest length.
* As a next step, from scientific python package we import fcluster. We use fcluster to form the clusters. The first parameter that we pass is the linkage type that stores the linkage heights value that is stored within this linkage type. Then we pass the 'criterion' as another parameter whose value is set to 'maxclust'. We also pass the maximum number of clusters. We choose number of clusters as 3. As a result of all these parameters the data will be split into 3 clusters.

Table

Description automatically generated

**Wards Linkage Method**

* We perform the same set of steps as we perfomed while creating Dendrogram using Average Linkage method.

Chart

Description automatically generated

* We truncate the Dendrogram using ‘lastp’ value as 10

Chart

Description automatically generated

# 

# We truncate the Dendrogram using ‘lastp’ value as 3 and linkage method as Wards Linkage method.

Chart, line chart

Description automatically generated

* On observing the dendrogram, we identify a cut-off point. We again choose the cut-off where the jump or vertical lines should be of highest length.
* from the above dendrogram, it is very evident that the distance after 10 is the highest. In a dendrogram greater the difference in height, more the dissimilarity within the objects i.e. heterogeneity between the groups. Hence optimum number of cluster is 3.
* We again use fcluster to form the clusters using the same parameters.

**Cluster Profiling**

Table

Description automatically generated

* Cluster 1 is having highest mean values for all the variables except ‘min\_payment\_amt’. The count of Cluster 1 customers is 70. Cluster 1 represents the ‘Best Customer’ category having high spending amount and high credit limit on an average.
* Cluster 2 is having the lower mean values for all the variables except ‘min\_payment\_amt’ and ‘max\_spent\_in\_single\_shopping’. The count of Cluster 2 customers is 67 and it represents the low income customers.
* Cluster 3 is having the medium mean values for all the variables except ‘min\_payment\_amt’ and ‘max\_spent\_in\_single\_shopping’ which is lower than Cluster 2. The count of Cluster 3 customers is 73 and it represents the middle class customers.

# Choosing 3 or 4 clusters seems like a good optimal number of clusters for the given data. We have grouped our data using 3 clusters only. The 3 clusters that we have chosen has divided our data into high, medium and low category.

**Chart, box and whisker chart

Description automatically generated**

# Chart, scatter chart Description automatically generated

# From the above 2 graphs i.e. Box Plot and Scatter Plot we conclude the following:

# Cluster 1 has higher values for all variables as compared to other Clusters. We can say that Cluster 1 represents the best customer segment or higher class customers. Cluster 1 shows lower value for the variable 'min\_payment\_amt'. We can see from the scatter plot that the customers having the highest spendings in Cluster 1 are the ones making highest advance payments and also having highest credit card limit. These are the customers who have spent maximum amount in a single shopping.

# Cluster 2 has the lowest values for all variables as compared to other Clusters except for the variables 'min\_payment\_amt' and 'max\_spent\_in\_single shopping'. We also observe that Cluster 2 has very low value for 'probability\_of\_full\_payment' and 'credit\_limit'. We can say that Cluster 2 represents low income customers and therefore less expenditure customer. The bank should be cautious before providing any loan to these customers. Customers in Cluster 2 are not having high spendings as compared to Cluster 1 and therefore they are having very low credit limit. These are the customers who make lowest advance payments. They have very high value for the variable ‘min\_payment\_amt’ i.e. average minimum amount paid by the customer while making payments for credit card bill purchases made monthly.

# Cluster 3 has intermediate values for all variables except 'min\_payment\_amt' and 'max\_spent\_in\_single\_shopping'. Cluster 3 represents middle class customers. The bank can provide some promotional advertisements or offers so that 'max\_spent\_in\_single\_shopping' can be increased. For customers in Cluster 3, highest spenders are the ones making highest advance payments but relatively less than Cluster 1 customers. They have high credit limit but can be higher as they are the potential customers who can move to best customer segment.

# 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.

**K-Means Clustering**

* K-Means clustering is a unsupervised learning algorithm whose goal is to find groups or assign data points to clusters on the basis of their similarity. This means that the points in same cluster are similar to each other and in different clusters dissimilar with each other.
* We imported from the sklearn python library (submodule cluster), the functions called as KMeans.
* K\_means technique is always applied on a scaled data. Differences in the scales across input variables may increase the difficulty of the problem being modelled. We have already scaled the data previously.
* We are building a K\_means model using the KMeans function and passing two parameters – ‘n\_clusters’ and ‘random\_state’. We first apply k\_means clustering technique with number of clusters as 1 and random state 1. Then we fit the scaled data into the KMeans object.
* From this we extract the first output as ‘labels\_’ which provides the cluster mapping. We can go ahead and map this or attach this as a column to the original dataframe. The second output that we are interested in is called the ‘inertia\_’ which is the total within sum of squares when k = 1. The within sum of squares for each cluster is computed by adding and squaring the distance between the centroid and its every observation.
* We perform the above mentioned step again but this time we set the value of ‘n\_clusters’ to 2. We will observe that as we increase the number of clusters for k\_means technique, the within the sum of square score keeps on decreasing. Larger the drop of within sum of squares as we keep increasing the k value, the better it is for the model. If the drop is not significant then it means additional cluster is not useful for the model.
* Now we will create an empty object which will store the inertia values for each iteration of the k\_means algorithm. At each iteration, we will create an object for KMeans model with k value running from 1 to 10. We will append the corresponding inertia for each model into out empty object. We will get the following set of values after running all the iterations:

Table

Description automatically generated

* Now we have within sum of squares from k=1 to k=10. This we can plot on a simple graph and see if there is a significant drop within the sum of squares value after we increase the number of clusters.

Chart, line chart

Description automatically generated

* One of the greatest challenges is to understand and know beforehand how many clusters we require as an output prior to running the k means algorithm itself.
* wss plot is within sum of squares plot. The within sum of squares for any cluster is the distance between the centroid and all observations which is squared and added up. This measure of wss is used to build a wss plot.
* In elbow plot we keep track of the wss value for a range of different values of k. We then look for a value where rate of reduction in WSS begins to decline. This signifies that adding an extra cluster is not obtaining enough clarity of cluster separation to justify further increasing k.
* we will track the plot where decrease in wss begins to flatten out compared to increasing k values as it is not revealing more signals.

**Table

Description automatically generated**

* This table shows the clusters to the dataset and also individual silhouette width score.
* We use the Silhouette score to find whether the mapping of each customer to a cluster is correct or not.
* The Silhouette score function computes the average of all the silhouette widths. The Silhouette sample function computes the silhouette widths for each and every row. These two modules are available in sklearn.metrics submodule.
* The input parameters required for running the silhouette\_score as well as silhouette\_sample are the scaled data and the label mapping.
* The silhouette score is 0.400805 which is on the positive side, we can say that this is a well distinguished set of clusters. So the 3 clusters that is created on an average have a silhouette score of 0.400805.
* for calculating the individual silhouette widths we can call the silhouette\_sample function. The input parameters are the same.

**Silhouette score using 3 clusters is 0.4008059221522216**

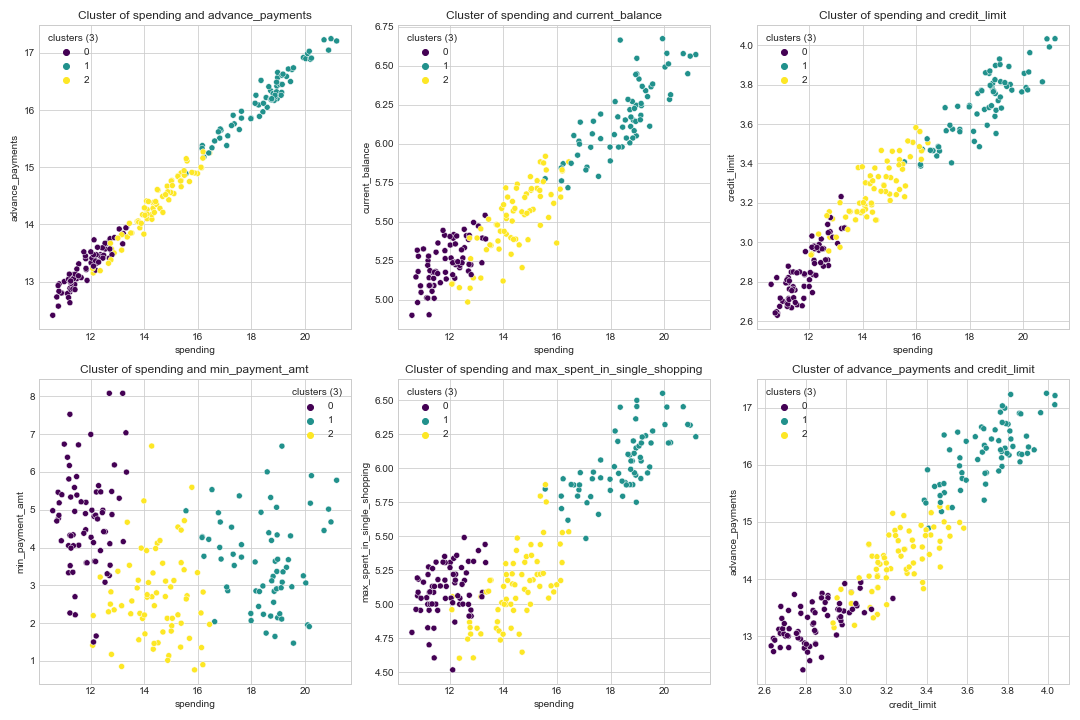
**Silhouette score using 3 clusters is 0.3373662527862716**

# **silhouette score is better for 3 clusters than for 4 clusters. So final clusters will be 3.**

**Cluster Profiling**

# 

* From the above table we can observe that the Cluster 0 is having the lowest values for the variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’. So we can conclude that Cluster 0 represents low spender customers or low income customers or low expenditure customers.
* Cluster 1 is having the highest values for variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’ except ‘min\_payment\_amt’. So we can conclude that Cluster 1 represents high spender customers or high income customers or high expenditure customers. This is the best customer segment.
* Cluster 2 is having medium values for variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’ except ‘min\_payment\_amt’. So we can conclude that Cluster 2 represents medium spender customers or medium income customers or medium expenditure customers.



* In the above graph we are Visualizing the clusters for each variable using scatter plots. We can see that the data is segregated into 3 clusters – Cluster 0, Cluster 1, Cluster 2 separating each other with distinct colours as purple, yellow and green respectively.
* Cluster 0 is having the lowest values for all the variables except for ‘min\_payment\_amt’. Customers in Cluster 0 are not having high spendings as compared to Cluster 1 and therefore they are having very low credit limit. These are the customers who make lowest advance payments. They have very high value for the variable ‘min\_payment\_amt’ i.e. average minimum amount paid by the customer while making payments for credit card bill purchases made monthly. We should send them more frequent reminders for their pending payments. We should provide the customers some attractive discount offers on early payments or timely payments so that they can be promoted to make the payments on time.

# For Cluster 1 we can see from the scatter plot that the customers having the highest spendings are the ones making highest advance payments and also having highest credit card limit. These are the customers who have spent maximum amount in a single shopping. We can have various offers for customers belonging to Cluster 1 like high reward points. We can have various promotional schemes for these customers like no cost EMI. The customers in Cluster 1 are having highest value for the variable ‘max\_spent\_in\_single\_shopping’. So we can offer these customers with high discounted offers on full payment for their purchase.

# For customers in Cluster 3, highest spenders are the ones making highest advance payments but relatively less than Cluster 1 customers. They have high credit limit but can be higher as they are the potential customers who can move to best customer segment. The customers in this segment should be informed about the advantages of being a premium member and should be informed about the benefits of the same. They should be promoted to make purchases from the luxurious brands and therefore they will increase their credit limit while managing their credit cards transactions because with proper management these customers can move into Cluster 1 segment. The customers should be informed about using their credit cards across a number of platforms as compared to what they are already using

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

* The Clusters that we have created are identified and described by their centroids or cluster centres. Cluster Profiling is the process of explaining in detail the properties of the cluster centroids so that the Business can benefit from the same.

**Hierarchical Clustering**

Table

Description automatically generated

# 

* Cluster 1 is having highest values for all the variables except ‘min\_payment\_amt’. The count of Cluster 1 customers is 70. Cluster 1 represents the ‘Best Customer’ category having high spending amount and credit high credit limit.
* Cluster 2 is having the lower values for all the variables except ‘min\_payment\_amt’ and ‘max\_spent\_in\_single\_shopping’. The count of Cluster 2 customers is 67 and it represents the low income customers.
* Cluster 3 is having the medium values for all the variables except ‘min\_payment\_amt’ and ‘max\_spent\_in\_single\_shopping’ which is lower than Cluster 2. The count of Cluster 3 customers is 73 and it represents the middle class customers.

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

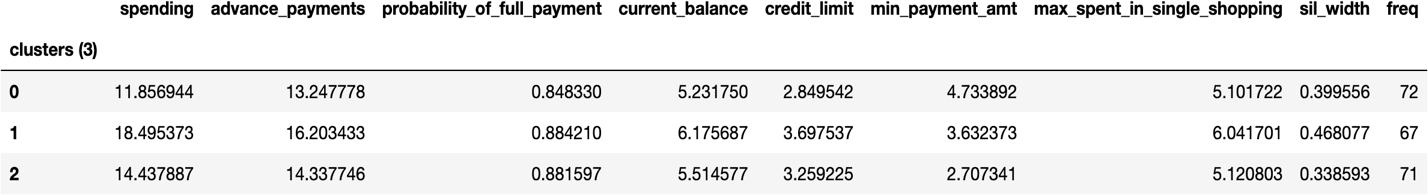
Description automatically generated

# Cluster 1 has higher values for all variables as compared to other Clusters. We can say that Cluster 1 represents the best customer segment or higher class customers. Cluster 1 shows lower value for the variable 'min\_payment\_amt'. Customers in Cluster 1 are premium high-net worth customers who make expensive purchases on their credit cards. We can have various offers for customers belonging to Cluster 1 like high reward points. We can have various promotional schemes for these customers like no cost EMI. The customers in Cluster 1 are having highest value for the variable ‘max\_spent\_in\_single\_shopping’. So we can offer these customers with high discounted offers on full payment for their purchase.

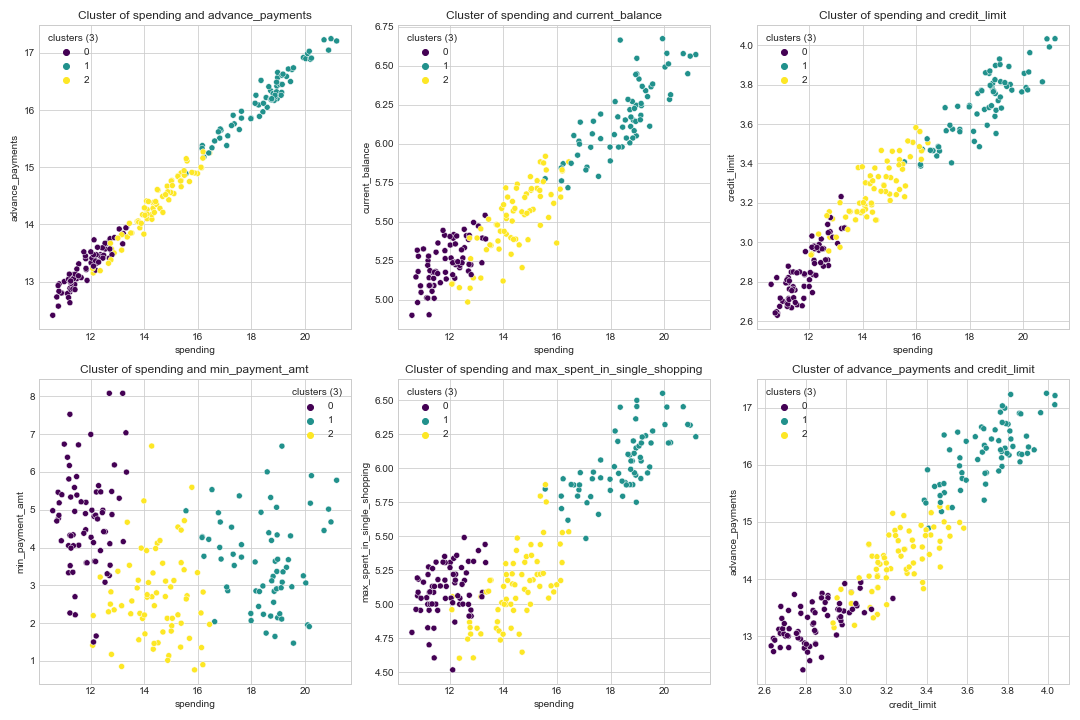
# Cluster 2 has the lowest values for all variables as compared to other Clusters except for the variables 'min\_payment\_amt' and 'max\_spent\_in\_single shopping'. We also observe that Cluster 2 has very low value for 'probability\_of\_full\_payment' and 'credit\_limit'. We can say that Cluster 2 represents low income customers and therefore less expenditure customer. The bank should be cautious before providing any loan to these customers. They should not provide high credit limit to these customers. These can be the customers who have recently bought credit cards or even can be the youths who have started working recently. These are the potential customers who can move into Cluster 2 or middle class customers. We should plan discount offers on the credit card usage on the basis of their purchasing capacity which should be different from customers belonging to Cluster 1. We should send them more frequent reminders for their pending payments. We should provide the customers some attractive discount offers on early payments or timely payments so that they can be promoted to make the payments on time.

# Cluster 3 has intermediate values for all variables except 'min\_payment\_amt' and 'max\_spent\_in\_single\_shopping'. Cluster 3 represents middle class customers. The bank can provide some promotional advertisements or offers so that 'max\_spent\_in\_single\_shopping' can be increased. These are the potential customers who can move to Cluster 1 or the premium high net worth customers. The customers in this segment should be informed about the advantages of being a premium member and should be informed about the benefits of the same. They should be promoted to make purchases from the luxury brand and therefore they will increase their credit limit while managing their credit cards transactions because with proper management these customers can move into Cluster 1 segment. The customers should be informed about using their credit cards across a number of platforms as compared to what they are already using.

**K-Means Clustering**



* From the above table we can observe that the Cluster 0 is having the lowest values for the variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’. So we can conclude that Cluster 0 represents low spender customers or low income customers or low expenditure customers.
* Cluster 1 is having the highest values for variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’ except ‘min\_payment\_amt’. So we can conclude that Cluster 1 represents high spender customers or high income customers or high expenditure customers. This is the best customer segment.
* Cluster 2 is having medium values for variables ‘spending’, ‘advance\_payments’, ‘probability\_of\_full\_payment’, ‘current\_balance’, ‘credit\_limit’ and ‘max\_spent\_in\_single\_shopping’ except ‘min\_payment\_amt’. So we can conclude that Cluster 2 represents medium spender customers or medium income customers or medium expenditure customers.



# Cluster 1 has higher values for all variables as compared to other Clusters. We can say that Cluster 1 represents the best customer segment or higher class customers. Cluster 1 shows lower value for the variable 'min\_payment\_amt'. Customers in Cluster 1 are premium high-net worth customers who make expensive purchases on their credit cards. We can have various offers for customers belonging to Cluster 1 like high reward points. We can have various promotional schemes for these customers like no cost EMI. The customers in Cluster 1 are having highest value for the variable ‘max\_spent\_in\_single\_shopping’. So we can offer these customers with high discounted offers on full payment for their purchase.

# Cluster 0 has the lowest values for all variables as compared to other Clusters except for the variables 'min\_payment\_amt' and 'max\_spent\_in\_single shopping'. We also observe that Cluster 0 has very low value for 'probability\_of\_full\_payment' and 'credit\_limit'. We can say that Cluster 0 represents low income customers and therefore less expenditure customer. The bank should be cautious before providing any loan to these customers. They should not provide high credit limit to these customers. These can be the customers who have recently bought credit cards or even can be the youths who have started working recently. These are the potential customers who can move into Cluster 0 or middle class customers. We should plan discount offers on the credit card usage on the basis of their purchasing capacity which should be different from customers belonging to Cluster 1. We should send them more frequent reminders for their pending payments. We should provide the customers some attractive discount offers on early payments or timely payments so that they can be promoted to make the payments on time.

# Cluster 2 has intermediate values for all variables except 'min\_payment\_amt' and 'max\_spent\_in\_single\_shopping'. Cluster 2 represents middle class customers. The bank can provide some promotional advertisements or offers so that 'max\_spent\_in\_single\_shopping' can be increased. These are the potential customers who can move to Cluster 1 or the premium high net worth customers. The customers in this segment should be informed about the advantages of being a premium member and should be informed about the benefits of the same. They should be promoted to make purchases from the luxurious brands and therefore they will increase their credit limit while managing their credit cards transactions because with proper management these customers can move into Cluster 1 segment. The customers should be informed about using their credit cards across a number of platforms as compared to what they are already using.

# **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).

# 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:

# 

* Shape of the data is (3000, 10) i.e. 10 columns and 3000 rows.
* The sample of the dataset i.e. the first five rows appear to be perfect.
* After checking and gathering more information about the data, we find that there are total 3000 entries in the dataset and all columns have 3000 non-null values.
* Age and Duration are of data type int64.
* Agency\_Code, Type, Claimed, Channel, Product Name and Destination are of data type object.
* Commission and Sales are of data type float64.
* The dataset has no null or missing values.
* The dataset has 139 duplicate values. We are not removing the duplicates as there are no unique identifiers given. so the duplicated data can belong to different customers as it is possible that a travel company can sell the same kind of tour package to similar demography.

# Exploratory Data Analysis

## Let us check the types of variables in the data frame.

# Column Non-Null Count Dtype

--- ------ -------------- ----- ---------

0 Age 3000 non-null int64

1 Agency\_Code 3000 non-null object

2 Type 3000 non-null object

3 Claimed 3000 non-null object

4 Commission 3000 non-null float64

5 Channel 3000 non-null object

6 Duration 3000 non-null int64

7 Sales 3000 non-null float64

8 Product Name 3000 non-null object

9 Destination 3000 non-null object

## Check for missing values in the dataset:

From the above results we can see that there are no missing values present in the dataset. There are total 3000 rows in the data set and all the columns have 3000 non-null values.

# 

## Describing the Data:

Table

Description automatically generated

* For Object data type variables Agency\_Code, Type, Claimed, Channel, Product Name and Destination, there are very less number of unique values.
* The topmost frequent value of:
  + - Agency\_Code is EPX with a frequency of 1365
    - Type is Travel Agency with a frequency of 1837
    - Claimed is No with a frequency of 2076
    - Channel is Online with a frequency of 2954
    - Product Name is Customised Plan with a frequency of 1136 ▪
    - Destination is ASIA with a frequency of 2465
* For the variable Age, the mean is 38.091 and median is 36.0. Since mean is larger than median, the variable ‘age’ is slightly skewed to the right. The minimum and maximum age of the customer is 8 and 84 respectively.
* The variable Commission is the commission received for tour insurance firm (in %). The mean is 14.529203 and the median is 4.63. Since the mean value is very high than the median, the variable ‘Commission’ is highly skewed to the right.
* The variable Duration is the duration of the tour (in days). The maximum and minimum duration of the tour is 4580 and -1 respectively. The value -1 is not possible as the number of days can’t be negative. Therefore we can say that it is an error value.
* The variable Sales is the amount worth of sales per customers in procuring tour insurance policies (in rupees in 100’s). The mean is 60.249913 and the median is 33.0. Since the mean value is very large as compared to median, the variable ‘Sales’ is highly skewed to the right.
* For the float and integers data type values like: Age, Commission, Duration and Sales the difference between its 75th percentile and Max value is very large, indicating there will be large number of outliers in the data.

## Univariate Analysis

## Box Plot

Chart

Description automatically generated

* The variable ‘Age’ is slightly skewed to the right. The median number of ‘Age’ is approximately around 36. The Inter-Quartile range is about 10 which means about 50% of ‘age’ is approximately between 31 and 41. The variable ‘Age’ has outliers.
* The variable ‘Commission’ is skewed to the right. The median number of ‘Commission’ is approximately around 5. The Inter-Quartile range is about 18 which means 50% of ‘Commission’ is approximately between 0 and 18. The variable ‘Commission’ has a lot of outliers.
* The variable ‘Duration’ is skewed to the right. The median number of ‘Duration’ is approximately around 4. The Inter-Quartile range is around 50 which means 50% of ‘Duration’ is approximately between 63 and 11. The variable ‘Duration’ has a lot of outliers.
* The variable ‘Sales’ is skewed to the right. The median number of ‘Sales’ is approximately around 30. The Inter-Quartile range is around 50 which means 50% of ‘Sales’ is approximately between 70 and 20. The variable ‘Sales’ has a lot of outliers.
* We prefer not to treat outliers here. Treating outliers results in models having better performance but the models lose out on generalization. Also, since the outliers do not directly impact any of the three models, treating the outliers is not necessary at this stage.

## Hist Plot

Chart, line chart

Description automatically generated

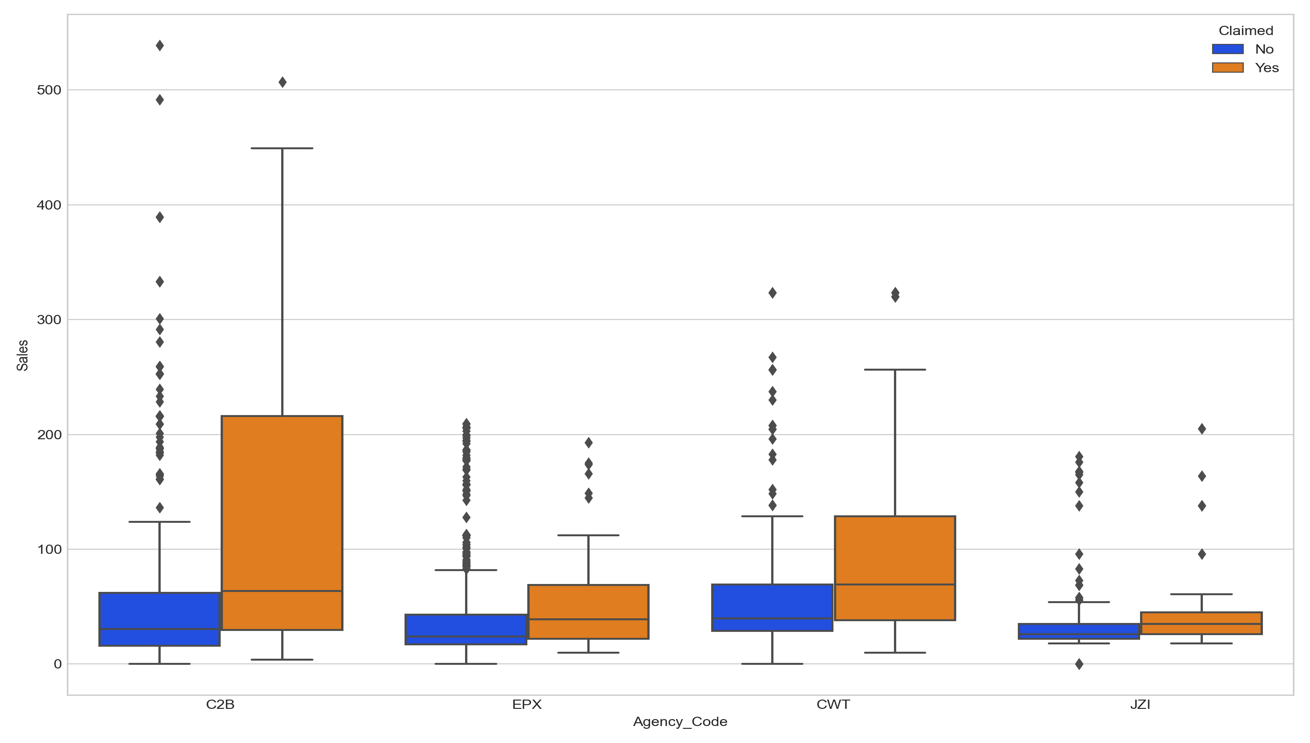
* For the variable ‘Age’ we can see that it is unimodal. This plot shows the distribution of data from 18 to 78. We can clearly see from the distribution plot that the variable is little skewed to the right but it is almost normally distributed. About 50% of ‘age’ is approximately between 31 and 41.
* For the variable ‘Commission’ we can see that it is unimodal. This plot shows the distribution of data from 0 to 210. We can clearly see from the distribution plot that the variable is highly skewed to the right. About 50% of ‘Commission’ is approximately between 0 and 18.
* For the variable ‘Duration’ we can see that it is unimodal. This plot shows the distribution of data from 0 to 4500 approximately. We can see from the distribution plot that the variable is highly skewed to the right. About 50% of ‘Duration’ is approximately between 63 and 11. The variable ‘Duration’ has a lot of outliers.
* For the variable ‘Sales’ we can see that it is unimodal. This plot shows the distribution of data from 0 to 550 approximately. We can clearly see from the distribution plot that the variable is highly skewed to the right. About 50% of ‘Sales’ is approximately between 70 and 20. The variable ‘Sales’ has a lot of outliers.

## Agency Code

Chart, bar chart

Description automatically generated

* From the above graph we can conclude that Agency\_Code EPX has the maximum frequency and JZI has the minimum frequency.



* The box plot shows the split of sales with different tour firms. The tour firm with code C2B has highest claimed sales as compared to any other agency. The tour firm JZI has lowest claimed sales. The tour firm CWT has highest unclaimed sales. Overall all the agency firms has outliers.

## Product Name

Chart, bar chart

Description automatically generated

* From the above graph we can conclude that the Customised Plan is the most liked plan by the customers. The Gold Plan is the least liked plan by the customers.

Chart, box and whisker chart

Description automatically generated

* The box plot shows the split of sales with different product which is further filtered by claimed status.
* We can see that there are outliers for all the products. More number of claims were made by customers who had gold plan and silver plan. But customised plan is the most liked plan by the customers.

## Type

Chart, bar chart

Description automatically generated

* From the above graph we can conclude that Travel Agency type insurance firm has more frequency as compared to Airlines type insurance firm.

Chart, box and whisker chart

Description automatically generated

* From the above graph we can conclude that Airline type tour firm made more claimed sales as compared to Travel Agency. Also, both the type of tour firm has outliers.

## Channel

Chart, bar chart

Description automatically generated

* From the above graph it is evident that Online channel has maximum frequency i.e. the most of the claims were made online.

Chart, box and whisker chart

Description automatically generated

* From the boxplot it is evident that more number of claims were made using Online channel. But we can also see that whatever frequency of claims that were made offline almost all the claims were successful.

## Destination

Chart, bar chart

Description automatically generated

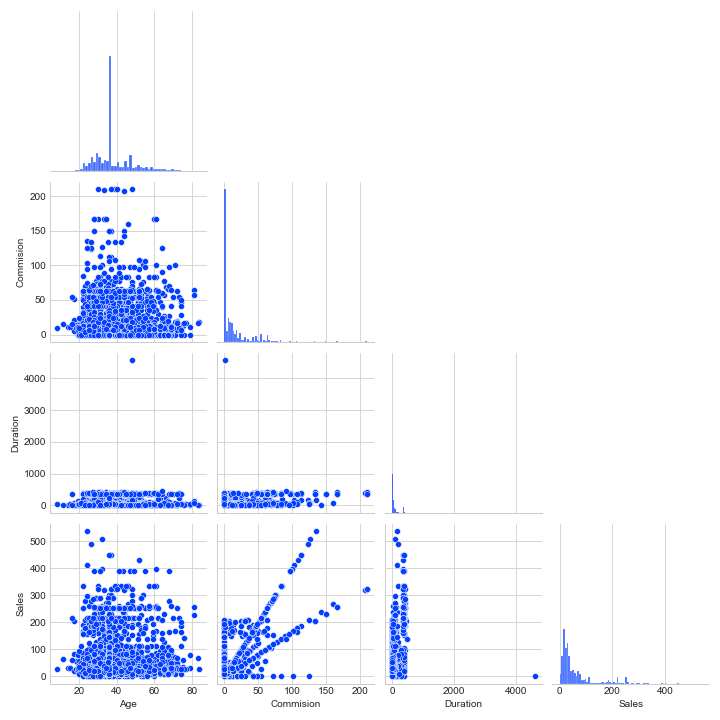
* From the above graph it is evident that Asia is the most preferred destination by travellers which is followed by America and then Europe.

Chart, box and whisker chart

Description automatically generated

* From the boxplot it is evident that more claims were made by customers who chose Asia as their destination. Less number of claims were made by the customers who chose Europe as their destination.

## Pairplot



* From the above Pairplot we can see that there is not much correlation between the variables. We cannot see any pattern in the scatter plots for any of the pair of variables and hence we can conclude that there is not much multicollinearity between the variables.
* Sales and Commission show some kind of positive relationship but it is not very strong.
* We will check the strength of above mentioned relationship via correlation matrix.

## Heatmap

A picture containing background pattern

Description automatically generated

* As the Pairplot suggested already that there is not much collinearity between variables but Sales and Commission did show some amount of positive relationship which is 0.77.
* We can conclude that higher the sales more the commissions received by the insurance agencies.

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

* CART models, RF models and ANN models in Python can take only numerical / categorical columns. It cannot take string / object types.
* The following code loops through each column and checks if the column type is object then converts those columns into categorical with each distinct value becoming a category or code.
* Some columns are integer data type and some columns are identified as object data type. The issue over here is that for building a Decision Tree model or CART model we have to ensure that there are no object data types. We should only have integer data types for both independent and dependent variable.
* The objective here is to transform all columns that are of object data types into integer data type.

# Column Non-Null Count Dtype

--- ------ -------------------- -------

0 Age 3000 non-null int64

1 Agency\_Code 3000 non-null int8

2 Type 3000 non-null int8

3 Claimed 3000 non-null int8

4 Commission 3000 non-null float64

5 Channel 3000 non-null int8

6 Duration 3000 non-null float64

7 Sales 3000 non-null float64

8 Product Name 3000 non-null int8

9 Destination 3000 non-null int8

* In the below table we can see the head of the dataset after datatype conversion

Table

Description automatically generated

Extracting the Target Column

* We will be extracting the target column 'Claimed' into separate vectors for training and test set.
* we have 3000 records in our data set. we will randomly split our data into 70 percent training and 30 percent testing.
* Before doing this we have to separate our data into independent variable and dependent variable separately.
* We will use the two functions to achieve the same - the drop function and the pop function.
* X will be the set of all independent variables. We will be dropping the 'Claimed' column from the dataset for X.
* Y will be the set of dependent variables.
* Now we will split the data into training and testing.
* from sklearn package and submodule model\_selection we will import the train\_test\_split function.
* we will call the train\_test\_split and we will pass the independent variable and dependent variable. The function train\_test\_split requires the dependent and independent variable separately and that is the reason why we have split into X and Y.
* The next parameter will be the test size. We will specify the proportion of test size as 0.3 or 30 percent.
* we will get four different outputs after executing the function which are - Training Independent variables, Testing Independent Variables, Training Dependent variables and Testing Dependent Variable.
* The dataset is split into ratio of 70:30, the training dataset is stored in X\_train and testing dataset in X\_test. The dimension of training and testing dataset is as below.
  + X\_train (2100, 9)
  + X\_test (900, 9)
  + train\_labels (2100,)
  + test\_labels (900,)

## Building A Decision Tree Classifier (CART)

* Decision Tree is a supervised approach which can be used either for classification / continuous regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It is a graphical representation of all possible solutions to a decision based on certain conditions.
* We are building a Decision Tree with some additional pruning parameters passed into the same. This is also called regularised decision tree model.
* We have chosen 'Gini' as the criterion. Gini Impurity is the mathematical measurement of how pure the information on the dataset is. This is a measurement of class uniformity.
* max\_depth is the length of the longest path from the tree root to a leaf. It should always be less than 20. but there is no particular threshold for max\_depth. It depends on model.
* min\_samples\_split specifies the minimum number of samples required to split an internal node should always be 1-10% of the entire record
* min\_samples\_leaf specifies the minimum number of samples required to be at a leaf node. It should always be 1/3rd of min\_samples\_split.
* For the given dataset we have taken the below sets of hyperparameters:
  + 'criterion': ['Gini']
  + 'max\_depth' : [7, 10, 13, 16]
  + 'min\_samples\_leaf': [50, 70, 80, 100]
  + 'min\_samples\_split' : [150, 200, 250, 300]
* GridSearchCV allows us to do Grid Search i.e. we can pass multiple input values for all the parameters and find out which combination of value will give the highest precision. After applying Grid Search validation, the best parameters were as below:
  + 'criterion': 'Gini'
  + 'max\_depth': 7
  + 'min\_samples\_leaf': 50,
  + 'min\_samples\_split': 150

Diagram

Description automatically generated

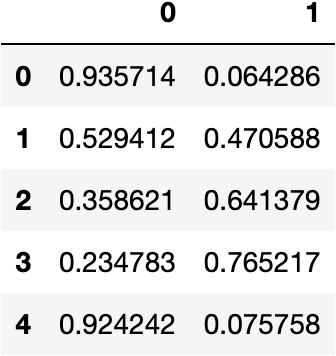
Feature Importance

* from the regularised Decision Tree model we will extract the single output called as feature\_importance\_.
* The output contains an array of values which indicates decrease in the node impurity weighted by the probability of reaching that node. It tells about the variables that is used in each and every split.

Chart, bar chart

Description automatically generated

* The variable which has a feature importance of 0 can be dropped from the dataset because that variable was never used in splitting the node within the Decision Tree. They will never be used for Training or Testing. The variable Channel is having 0 feature importance value and hence it can be dropped from the Dataset.
* We can say that the variable ‘Agency\_Code’, ‘Product Name’, ‘Sales’, ‘Commission’ are very important for splitting the node within the Decision Tree.
* The Final step involves predicting on the testing data using independent test variables and getting the predicted classes and Probabilities.

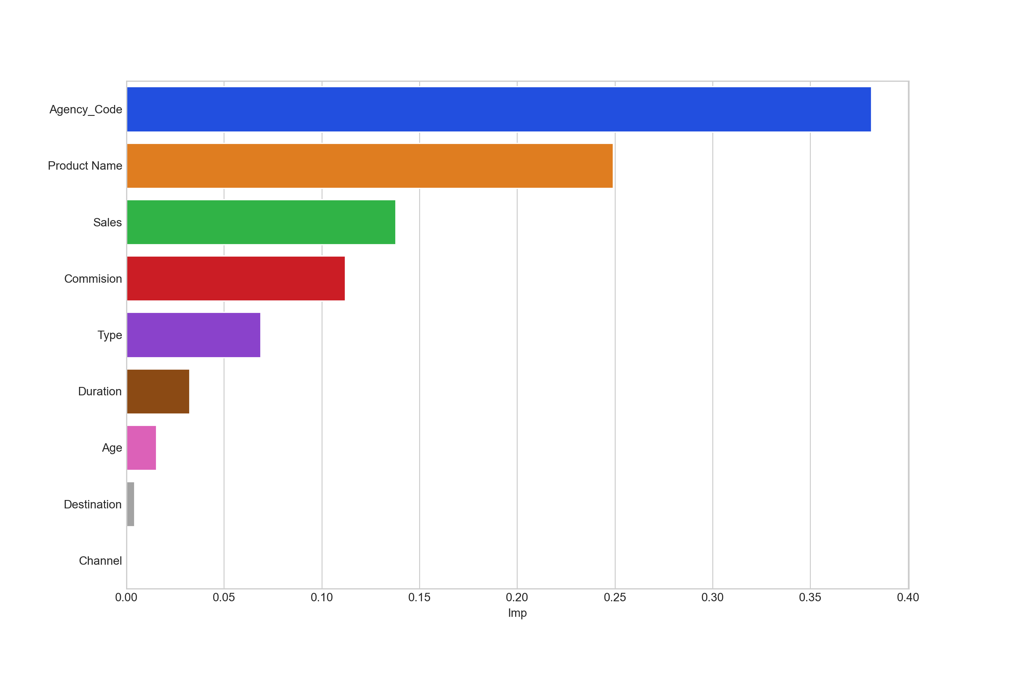


Building A Random Forest Classifier

* Random Forests is an ensemble machine learning technique that combines several base models in order to produce one optimal predictive model. Random Forests are a collection of decision trees. In random forests, each tree in the ensemble is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set.
* Random Forest has the ability to greatly increase the performance of the model based on expanding ideas from Decision Trees.
* This is also called ensemble technique since they rely on ensemble of models or multiple Decision Trees.
* max\_features states that out of the total number of independent features how many features should the random forest classifier use for evaluating and splitting the decision node in all decision trees. It will be less than or equal to number of independent variables.
* n\_estimators will be 1-5% of the entire record. It is the number of trees that we want to build within the random forest classifier.
* max\_depth is the length of the longest path from the tree root to a leaf. It should always be less than 20. but there is no particular threshold for max\_depth. It depends on model.
* min\_samples\_split specifies the minimum number of samples required to split an internal node should always be 1-10% of the entire record.
* min\_samples\_leaf specifies the minimum number of samples required to be at a leaf node. It should always be 1/3rd of min\_samples\_split.
* For the given dataset we have taken the below set of hyperparameters:
  + 'max\_depth': [7, 10, 13]
  + 'max\_features': [4, 5, 6]
  + 'min\_samples\_leaf': [70, 80, 100]
  + 'min\_samples\_split' : [200, 250, 300]
  + 'n\_estimators' : [200, 250, 300]
* GridSearchCV allows us to do Grid Search i.e. we can pass multiple input values for all the parameters and find out which combination of value will give the highest precision.
* we have created a dictionary called as 'param\_grid\_rfcl' which will be passed to the Grid Search function as a parameter. After applying Grid Search validation, the best parameters were as below:
  + 'max\_depth': 7
  + 'max\_features': 4
  + 'min\_samples\_leaf': 70
  + 'min\_samples\_split': 200
  + 'n\_estimators': 300

Feature Importance

* from the model we will extract the single output called as feature\_importance\_.
* The output contains an array of values which indicates decrease in the node impurity weighted by the probability of reaching that node. It tells about the variables that is used in each and every split.



* The variable which has a feature importance of 0 can be dropped from the dataset because that variable was never used in splitting the node within the Decision Tree. They will never be used for Training or Testing.
* We can say that the variable ‘Agency\_Code’, ‘Product Name’, ‘Sales’, ‘Commission’ are very important for splitting the node within the Random Forest Tree.

Artificial Neural Network (ANN)

* It is a machine learning algorithm that is roughly modelled around what is currently known about the human brain. This is also called as Black Box technique.
* The various applications of ANN are as follows:
  + Image and voice recognition
  + To find complex patterns amongst very large datasets
  + Email engine suggesting sentence completion
  + Machine translating from one language to another
* For ANN (Artificial Neural Networks), it is important that we scale the data.
* This is done so that each variable holds the same weightage in the output so that the model is not affected by one variable by giving it more weightage.
* Scaling will convert all the data into the same scale range.
* We have already split the data into train and test using the train\_test\_split function. We will scale the 'X\_train' and 'X\_test' data and assign it to 'x\_train' and 'x\_test' variable. We will perform fit and transform on the 'X\_train' data. The training data is fitted into the Standard Scaler and then transformed. For the 'X\_test' data we will only transform this is because test data has to be scaled with respect to the scaling properties of training data.
* hidden layer is the layer between the input layer and the output layer hidden layer sizes is passed as a tuple of information. The size of tuple will indicate the number of hidden layers and each value mentioned in the tuple indicates the size of neurons in each hidden layers.
* max\_iter means the model is not allowed to run more than the specified number mentioned for this parameter. In each iteration the random synaptic weights that are initiated will be updated.
* solver is the argument to set the optimization algorithm here. It is used for weight optimization.
* tol is the threshold level. It is for the optimization. When the loss or score is not improving by the least tol for n\_iter\_no\_change consecutive iterations, the training stops.
* Activation function is a mechanism by which the artificial neuron processes incoming information and passes it throughout the network. They basically decide whether the neuron should be activated or not. By default we are using 'Relu' or Rectified Linear Activation Function for our model.
* For the given dataset we have taken the below sets of hyperparameters:
  + 'hidden\_layer\_sizes' : [50, 100, 200]
  + 'max\_iter' : [2500, 3000, 4000]
  + 'solver' : ['adam']
  + 'tol' : [0.01]
  + ‘activation’ : Relu
* GridSearchCV allows us to do Grid Search i.e. we can pass multiple input values for all the parameters and find out which combination of value will give the highest precision. After applying Grid Search validation, the best parameters were as below:
  + 'hidden\_layer\_sizes': 200
  + 'max\_iter': 2500
  + 'solver': 'adam'
  + 'tol': 0.01
* Using the GridSearchCV first the best params are identified and MLP is created using those params with training data.

Text

Description automatically generated

Getting the Predicted classes and Probabilities

* The Final step involves predicting on the testing data using independent test variables and getting the predicted classes and Probabilities.

Table

Description automatically generated

# **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

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

**A picture containing treemap chart

Description automatically generated**

**Confusion Matrix – Testing Data**

**Treemap chart

Description automatically generated with medium confidence**

* In our dataset for the target variable 'Claimed', 0 means the person has not claimed and 1 means the person has claimed.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the person actually made a claim but the model predicted that the person will not. It will be a more costlier mistake for the management. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the person has made a claim and the model also predicted the same.
* From the above table we can see that for training data FN = 271 and TP = 358 and testing data FN = 168 and TP = 127.
* The performance metric which considers FN and TP as the measuring parameter is Recall or Sensitivity. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* 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.

**Classification Report – Training Data**

**Table

Description automatically generated**

**Classification Report – Testing Data**

**Table

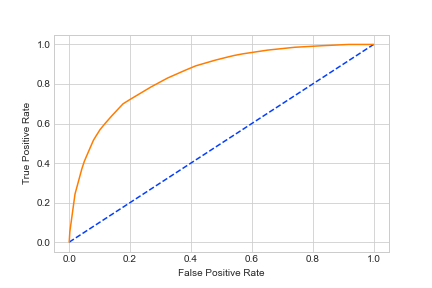
Description automatically generated**

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 80% and Accuracy for testing data is 74.5%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Recall for training data is 56.9% and Recall for testing data is 43%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 70.6% and precision for testing data is 67.5 %.

**AUC and ROC Curve**

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is. It is a graph showing the performance of a classification model at all classification thresholds.

**AUC and ROC Curve – Training Data**

****

**AUC and ROC Curve – Testing Data**

**Chart, line chart

Description automatically generated**

* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP)
* The above graph is constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.7.
* For the training data ROC graph if we take 0.2 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is slightly larger than 0.7.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 84.5% and for testing data it is 79.8%.

Random Forest

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

A picture containing square

Description automatically generated

**Confusion Matrix – Testing Data**

**Chart, treemap chart

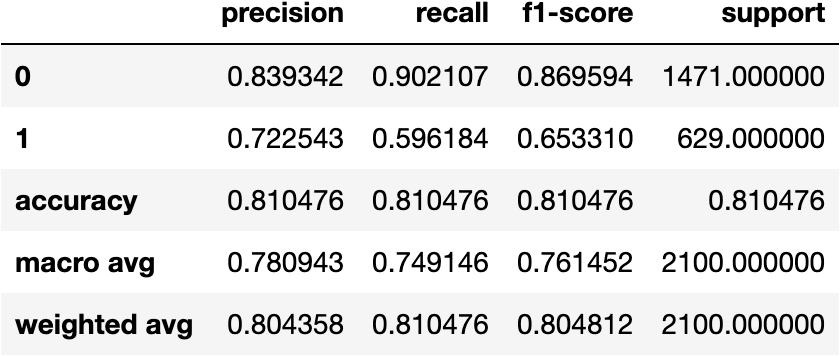
Description automatically generated**

* In our dataset for the target variable 'Claimed', 0 means the person has not claimed and 1 means the person has claimed.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the person actually made a claim but the model predicted that the person will not. It will be a more costlier mistake for the management. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the person has made a claim and the model also predicted the same.
* From the above table we can see that for training data FN = 254 and TP = 375 and testing data FN = 153 and TP = 142.
* The performance metric which considers FN and TP as the measuring parameter is Recall or Sensitivity. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* 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.

**Classification Report – Training Data**

****

**Classification Report – Testing Data**

**Table

Description automatically generated**

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 81.04% and Accuracy for testing data is 77.11%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Recall for training data is 59.61% and Recall for testing data is 48.13%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 72.25% and precision for testing data is 72.82%.

**AUC and ROC Curve**

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is. It is a graph showing the performance of a classification model at all classification thresholds.

**AUC and ROC Curve – Training Data**

Chart, line chart

Description automatically generated

**AUC and ROC Curve – Testing Data**

Chart, line chart

Description automatically generated

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is.
* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP) The above graph is constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.6 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 0.9.
* For the training data ROC graph if we take 0.6 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 1.0.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 84.5% and for testing data it is 79.8%.

Artificial Neural Networks

**Confusion Matrix**

* A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. It is used to determine the performance of the classification models for a given set of test data.

**Confusion Matrix – Training Data**

Chart, treemap chart

Description automatically generated

**Confusion Matrix – Testing Data**

Treemap chart

Description automatically generated with low confidence

* In our dataset for the target variable 'Claimed', 0 means the person has not claimed and 1 means the person has claimed.
* if we study the confusion matrix, we will find that False Negative i.e. FN is the most important for us because it signifies the condition when the person actually made a claim but the model predicted that the person will not. It will be a more costlier mistake for the management. This is the type 2 error.
* the second most important metric for us is True Positive i.e. TP because it signifies the condition when the person has made a claim and the model also predicted the same.
* From the above table we can see that for training data FN = 311 and TP = 318 and testing data FN = 167 and TP = 128. The performance metric which considers FN and TP as the measuring parameter is Recall or Sensitivity. FN is the Type 2 error and for Type 2 error we consider Recall or Sensitivity as the optimised performance measure for our problem statement.

**Classification Report**

* 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.

**Classification Report – Training Data**

Table

Description automatically generated

**Classification Report – Testing Data**

Table

Description automatically generated

* Accuracy of the model is the correct predictions divided by total number of observations. It suggests how accurately does the model classify the data points. Accuracy for training data is 77.57% and Accuracy for testing data is 76%. Higher the Accuracy of the model, stronger the prediction lower the accuracy weaker the prediction.
* The Recall of the model tells us how many of the true data points are identified as the true data points by the model. It describes the Type 2 error. Recall for training data is 50.55% and Recall for testing data is 43.38%.
* The precision of the model tell us among the data points identified as Positive by the model, how many are really positive. It describes the Type 1 error. precision for training data is 66.52% and precision for testing data is 72.63%.

**AUC and ROC Curve**

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is. It is a graph showing the performance of a classification model at all classification thresholds.

**AUC and ROC Curve – Training Data**

Chart, line chart

Description automatically generated

**AUC and ROC Curve – Testing Data**

**Chart, line chart

Description automatically generated**

* ROC curve is the technique for visualising the output of classification models to find out how good the performance is.
* It is the graph that is calculated between the true positive rate and false positive rate in the confusion matrix.
* True positive rate = TP/total positive - y axis
* False positive rate = FP/total negative - x axis
* ROC graph is drawn to show a trade-off between the benefits (TP) and costs (FP). The above graph has been constructed between 0 and 1 in x axis and 0 and 1 in y axis. This represents the percentage or the probabilities.
* The x axis is the negative scenario and y axis is the positive scenario. We expect a hight true positive rate and low false positive rate.
* For the testing data ROC graph if we take 0.6 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is slightly greater than 0.9.
* For the training data ROC graph if we take 0.6 as the cut-off to separate or classify the positives and the negatives the corresponding true positive rate is around 1.0.
* We know that the steeper the ROC curve, the stronger the model and the flatter the ROC curve , the weaker the model.
* AUC means Area under the ROC Curve. Larger the AUC, better the model because the more steeper it will be. The AUC for training data is 84.5% and for testing data it is 79.8%.

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

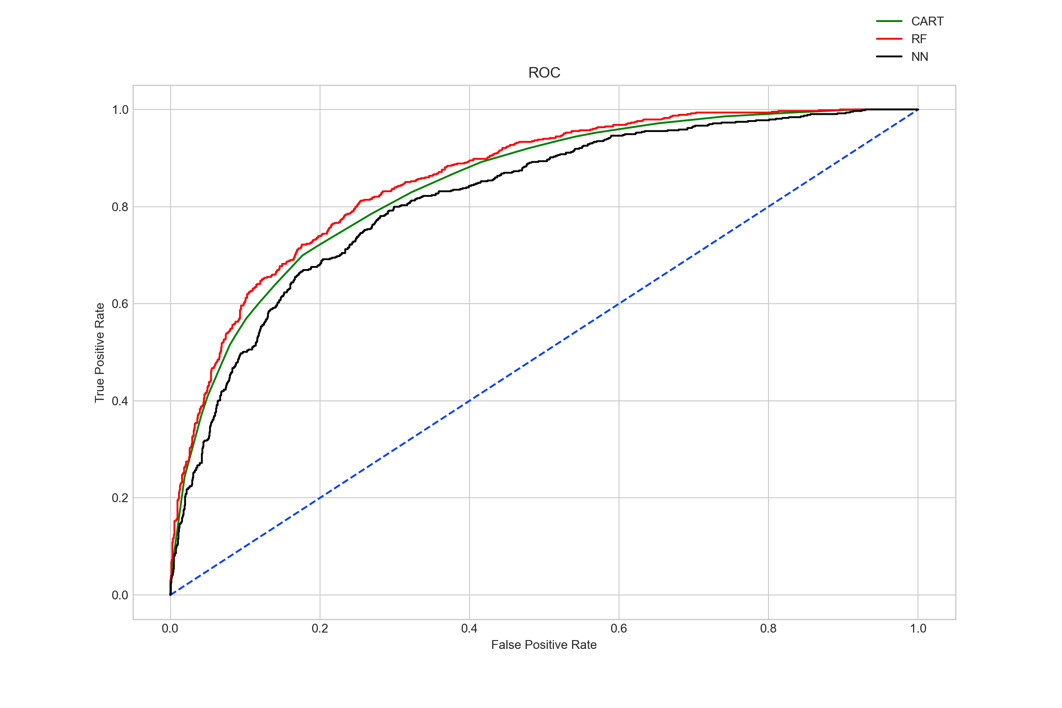
**Comparison of all performance metrics from 3 models:**

**Table

Description automatically generated**

* From the above table, we can say that none of the models have been Overfitted or underfitted.
* We could see that the Random Forest model has highest Accuracy on the Training data (0.81) as well as Testing data (0.7711).
* We can also conclude that we have lower Accuracy on the Artificial Neural Networks for Training data (0.77) as well as Testing Data (0.76).
* CART provides an accuracy for Training data (0.80) and for Testing data (0.745) which is relatively better than neural networks for Training Data.
* The F1 score and Precision for Random Forest for training data is (0.65) and (0.72) respectively and for Testing data is (0.58) and (0.73) respectively. The F1 score and Precision for Neural Network for training data is (0.57) and (0.67) respectively and for testing data is (0.54) and (0.72) respectively. The F1 score and precision for CART for Training data is (0.63) and (0.71) and for Testing data is (0.53) and (0.68) respectively. Therefore we can conclude that Random Forest performs better for Training data as well as Testing data as compared CART and Neural Network.

**AUC and ROC Curve – Training Data**



**AUC and ROC Curve – Testing Data**

**Chart, line chart

Description automatically generated**

* From the Graph we can also conclude that we have much higher AUC score and a steeper curve with Random Forest model for both the Training data (0.8588) and Test data (0.82) as compared to CART and Neural Network model.
* Out of the 3 models, Random Forest has slightly better performance than the CART and ANN model. Random Forest is performing the best as compared to CART and ANN and therefore we will use Random forest for more accurate predictions.
* Overall all the 3 models are reasonably stable enough to be used for making any future predictions. From CART and Random Forest Model, the variable Agency\_Code is found to be the most useful feature amongst all other features for predicting if a customer has claimed the insurance or not. The same can be observed from the Feature Importance table. All the three models are having high accuracy, AUC, Recall and precision values.
* All the models yields good results for the given problem, but we select the one that yields the best results and is well optimized for the Business Problem given to us. Therefore we select Random Forest as for highly accurate and precise predictions.

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

For the Business problem specified to us, we have built 3 different models – CART, Random Forest and Artificial Neural Network for the same. These models will predict the claim status for the company because this firm which is providing tour insurance is facing higher claim frequency. These models are built on training dataset and evaluated on testing dataset.

From the Feature Importance Tables for all the models we conclude the following:

* The variable which has a feature importance of 0 can be dropped from the dataset because that variable was never used in splitting the node within the Decision Tree. They will never be used for Training or Testing. The variable Channel is having 0 feature importance value and hence it can be dropped from the Dataset.
* We can say that the variable ‘Agency\_Code’, ‘Product Name’ and ‘Sales’ are very important for splitting the node for all the 3 models.

Since Agency\_Code is an important factor for our model Lets visualise the same using Bar Graph and Box plot.

Bar Graph

Chart, bar chart

Description automatically generated

* From the above graph we can conclude that Agency\_Code EPX has the maximum frequency and JZI has the minimum frequency. This means that maximum customers use Tour Firm EPX. Some training needs to be provided to the agency JZI so that they can increase their sales.

Box Plot

Chart, box and whisker chart

Description automatically generated

* The box plot shows the split of sales with different tour firms. The tour firm with code C2B has highest claimed sales as compared to any other agency. The tour firm JZI has lowest claimed sales. The tour firm CWT has highest unclaimed sales. Overall all the agency firms has outliers.
* we can see that the maximum claims is from agency code ‘C2B’ and surpasses all the other three agencies, hence the company needs to investigate why this tour firm has been having most highest claim frequency which may be due to several factors like injury, theft, lost baggage, some kind of accident that may contribute for the customer to seek higher claims.
* Some training needs to be provided to the agency JZI so that they can increase their sales. We can achieve this by providing some promotional events or run marketing campaign for the same.

‘Product Name’ is also an important feature for the model lets visualise the same using Bar Graph and Box plot.

Bar Graph

Chart, bar chart

Description automatically generated

* From the above graph we can conclude that the Customised Plan is the most liked plan by the customers. The Gold Plan is the least liked plan by the customers. The reason for the same might be its expansiveness.

Box Plot

Chart, box and whisker chart

Description automatically generated

* The box plot shows the split of sales with different product which is further filtered by claimed status.
* We can see that there are outliers for all the products. More number of claims were made by customers who had gold plan and silver plan. But customised plan is the most liked plan by the customers. As per the ML models, the Insurance Product type held highest importance for claims hence from below graph we can see that the Gold plan has the highest claims followed by Silver plan and Bronze plan of the tour insurance products. This might be due to the reason that Gold Plan covers a lot more aspects and areas which is relevant to an individual making the claims for the same. All other plans might not give so much flexibility to the customer for making claims.

‘Sales’ is also an important feature which contributes to the model prediction and therefore lets visualise the same using Distribution Plot

Chart, histogram

Description automatically generated

* From the above graph we can see the claim is maximum for the sale value in the range between 0 and 100 units($)
* Hence we can say that the maximum claim Is for the customers who belong to the low sales segment. As the sales segment is increasing the sales per customer in procuring tour insurance policies is decreasing.

The insurance products does not state/explain the coverage and disclaimers hence the products can be updated with new features and changes should be made for the same by getting feedback from customers on the reasons why claims have been highest for products like Gold Plan, Bronze Plan and Silver Plan. The reasons can be medical emergency like an accident, victim of crime, theft of luggage and cash etc. The company should get feedback on the customers profile for the three products to see why these three have higher claims.

Next the company needs to investigate or audit why the tour firm C2B has the highest claims. It can be due to various factors which might be internal or external. The insurance company can audit the firm or reach out to customers who buy the insurance from this agency to determine the reasons for the claims.

The company can next look at the low sales category and the reason for the claims why the low travel budget customers have been seeking claims which may be due to crime, damages , theft etc in these travel destinations or it can be due to other factors.

The ‘Channel’ variable is having 0 feature Importance value and thus does not contribute much to the model and therefore can be removed from the data. But there is on interesting inference that we can make from this column. Let’s visualise the same using a Box Plot.

Box Plot

Chart, box and whisker chart

Description automatically generated

* From the boxplot it is evident that more number of claims were made using Online channel. But we can also see that whatever frequency of claims that were made offline almost all the claims were successful.

The Firm should get behind this and find out the reason for the same.