# DATA MINING

Report by
Souravi

# Table of Contents

# INTRODUCTION

This report includes a detailed explanation of the approach taken, inferences, and insights addressing all two problems. It includes outputs such as graphs, tables, and all other relevant information. This Report does not include any codes.

## Cases Covered

**CLUSTERING**

**CART-RF-ANN**

# PROBLEM - 1

# SUMMARY

## BANK MARKETTING ANALYSIS

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

## Exploratory Data Analysis

### Data Description

### Data Dictionary for Market Segmentation:

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)

# PROBLEM 1

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

### Sample of the dataset

| spending | advance_payments | probability_of_full_payment | current_balance | credit_limit | min_payment_amt | max_spent_in_single_shopping |
|---|---|---|---|---|---|---|
| 19.94 | 16.92 | 0.875200 | 6.675 | 3.763 | 3.252 | 6.550 |
| 15.99 | 14.89 | 0.906400 | 5.363 | 3.582 | 3.336 | 5.144 |
| 18.95 | 16.42 | 0.882900 | 6.248 | 3.755 | 3.368 | 6.148 |
| 10.83 | 12.96 | 0.810588 | 5.278 | 2.641 | 5.182 | 5.185 |
| 17.99 | 15.86 | 0.899200 | 5.890 | 3.694 | 2.068 | 5.837 |

Table No. 1

### Data types of different variable

| Field Names | Non-Null Count | Dtype |
|---|---|---|
| spending | 210 non-null | float64 |
| advance_payments | 210 non-null | float64 |
| probability_of_full_payment | 210 non-null | float64 |
| current_balance | 210 non-null | float64 |
| credit_limit | 210 non-null | float64 |
| min_payment_amt | 210 non-null | float64 |
| max_spent_in_single_shopping | 210 non-null | float64 |

Table No. 2

### Observations¶

1. Dataset has a total of 210 rows and 7 columns.
2. All the fields are of float types.
3. There is no missing values.

## Checking for duplicates

There is No duplicates present in the dataframe.

## Missing data analysis for the variables

No missing values present in the data frame.

| Fields | Null Present |
|---|---|
| spending | FALSE |
| advance_payments | FALSE |
| probability_of_full_payment | FALSE |
| current_balance | FALSE |
| credit_limit | FALSE |
| min_payment_amt | FALSE |
| max_spent_in_single_shopping | FALSE |

Table No. 3

## Univariate Analysis

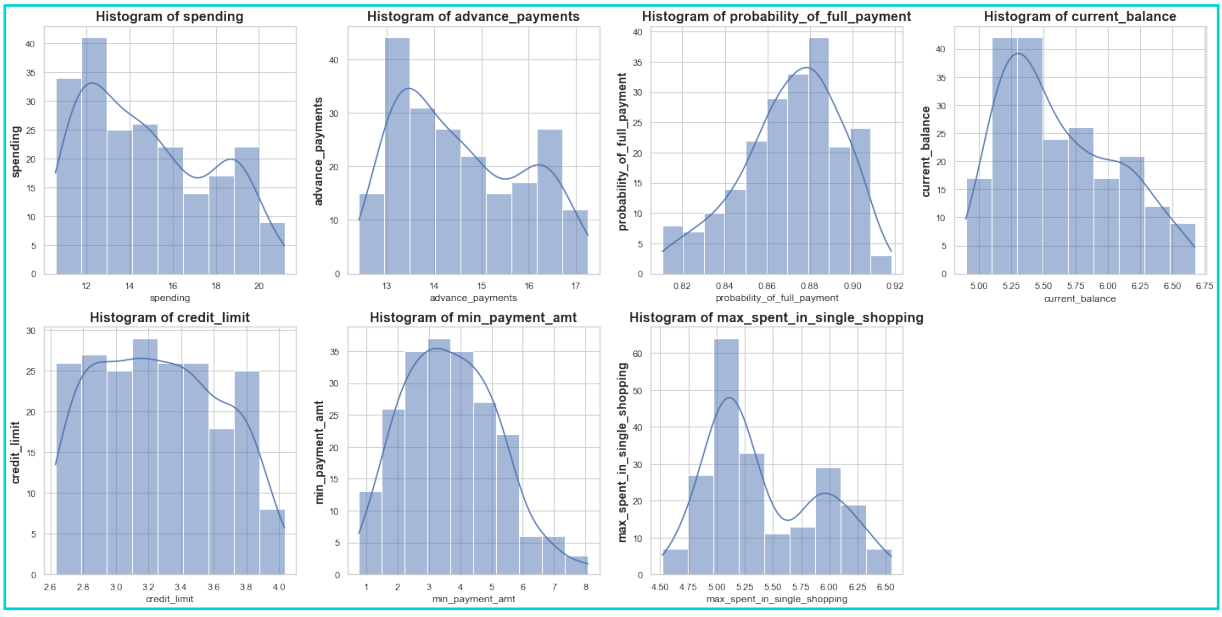



Figure No. 1

## Observations

### Positively Skewed

- Skewness of current_balance is 0.5217206481959235
- Skewness of min_payment_amt is 0.35742468670751903

### Bimodal Distribution

- Skewness of advance_payments is 0.38380604212562563
- Skewness of spending is 0.39702715402072153
- Skewness of max_spent_in_single_shopping is 0.5578758322317954

### Negatively Skewed

- Skewness of probability_of_full_payment is -0.5190516297742105

### Plateau or Multimodal Distribution

- Skewness of credit_limit is 0.13341648969738135

### Bivariant Analysis¶

This can be done both by Correlation Heat Map and PairPlot

### HeatMap

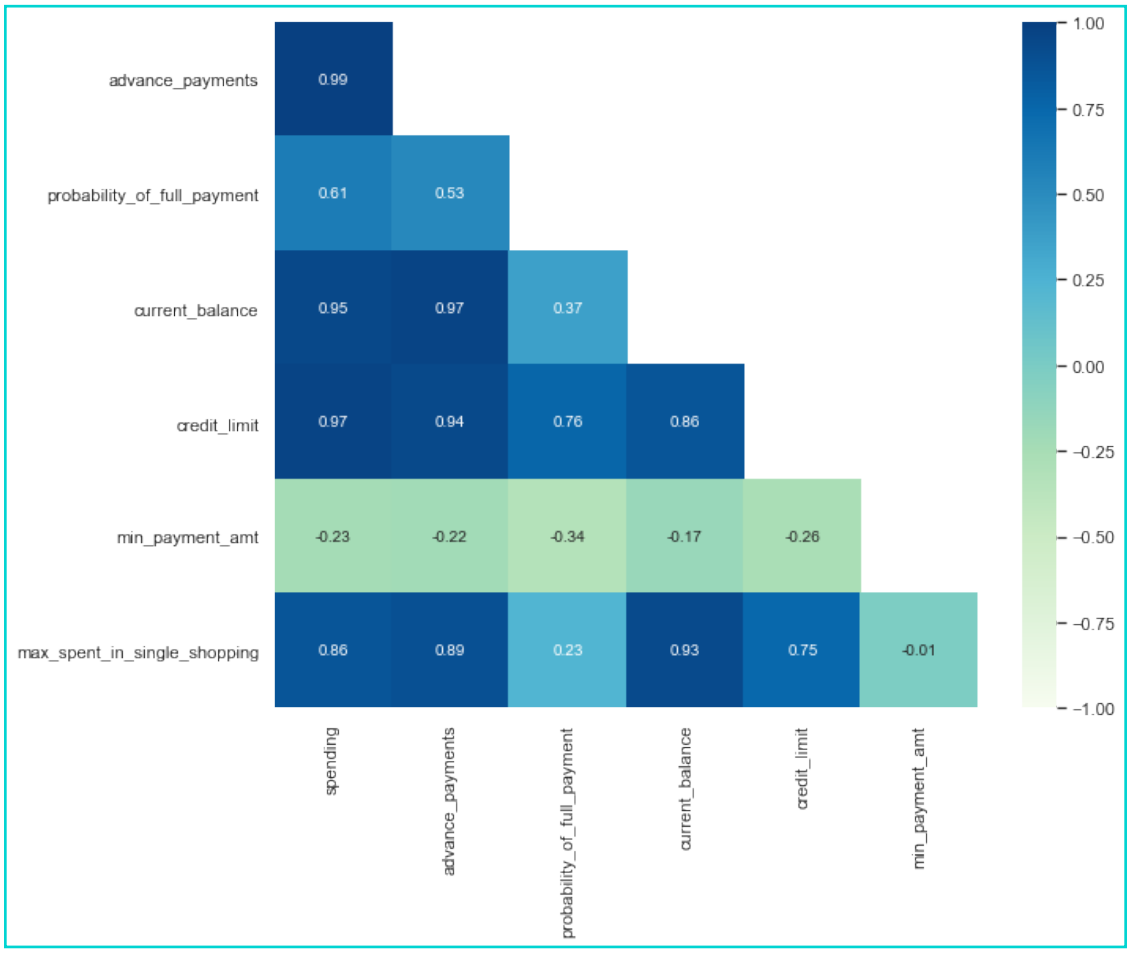



Figure No. 2

## Observations

1. Strong positive correlation can be observed in spending-advance_payments, spending-credit_limit , spending-current_balance, current_balance-max_spent_in_single_shopping.
2. No correlation is observed in fields like probability_of_full_payment-max_spent_in_single_shopping, min_payment_amt-max_spent_in_single_shopping etc.
3. No strong negative correlations.

## PairPlot

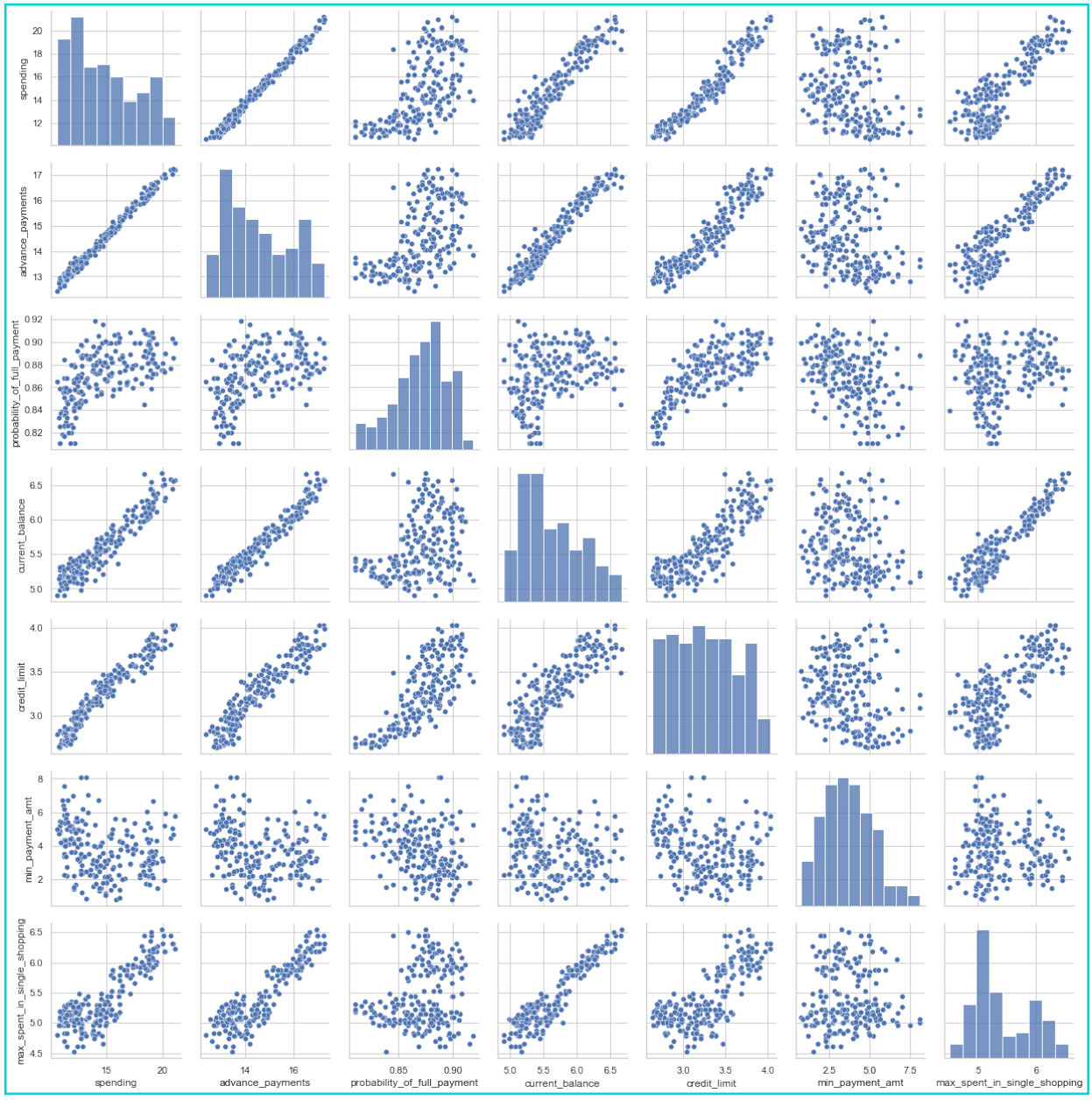



Figure No. 3

## Observations:¶

1. Positive covariance can be observed in spending-advance_payments, current_balance-spending, spending-credit_limit etc.
2. Negative Covariance can be observed in probability_of_full_payment-min_payment_amt etc.
3. No covariance seen in max_spent_in_single_shopping-probability_of_full_payment, probability_of_full_payment-credit_limit etc.

## BoxPlot

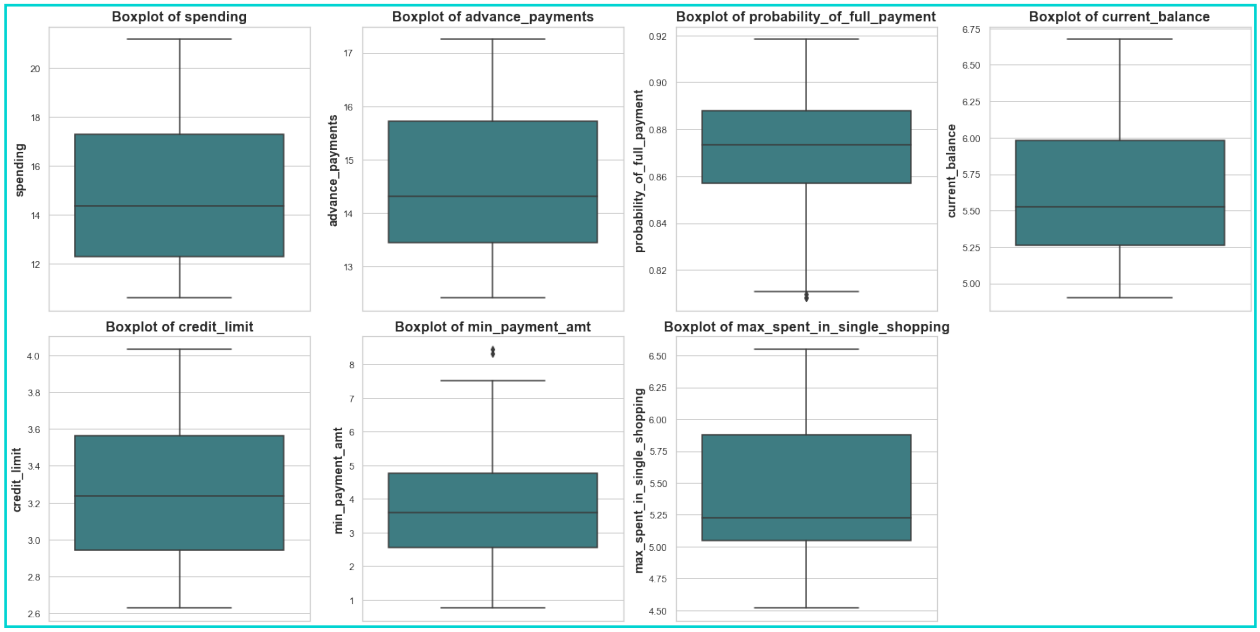



Figure No. 4

## Observations:¶

1. Outliers are present for probability_of_full_payment and min_payment_amt.
2. Outlier treatment can be done for the mentioned fields.

After treating Outliers, we observe the below Boxplots.

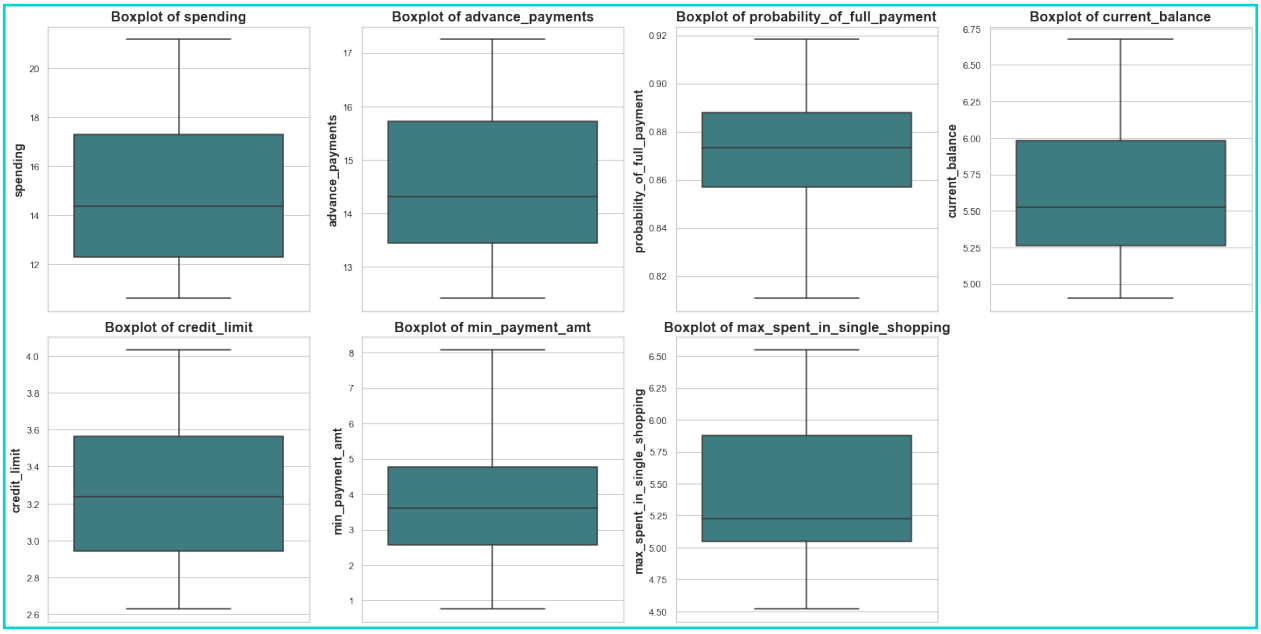

Boxplot of spending
spending
Boxplot of advance_payments
advance_payments
Boxplot of probability_of_full_payment
probability_of_full_payment
Boxplot of current_balance
current_balance
Boxplot of credit_limit
credit_limit
Boxplot of min_payment_amt
min_payment_amt
Boxplot of max_spent_in_single_shopping
max_spent_in_single_shopping

Figure No. 5

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

To determine if Scaling is required or not, we need to find the summary of all the fields and check if all those are in approximately same range.

### Descriptive Statistic

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **spending** | 210.0 | 14.8475 | 2.9096 | 10.5900 | 12.270 | 14.355 | 17.3050 | 21.1800 |
| **advance_payments** | 210.0 | 14.5592 | 1.3059 | 12.4100 | 13.450 | 14.320 | 15.7150 | 17.2500 |
| **probability_of_full_payment** | 210.0 | 0.87102 | 0.0235 | 0.81058 | 0.8569 | 0.8734 | 0.88777 | 0.91830 |
| **current_balance** | 210.0 | 5.62853 | 0.4430 | 4.89900 | 5.2622 | 5.5235 | 5.97975 | 6.67500 |
| **credit_limit** | 210.0 | 3.25860 | 0.3777 | 2.63000 | 2.9440 | 3.2370 | 3.56175 | 4.03300 |
| **min_payment_amt** | 210.0 | 3.69728 | 1.4946 | 0.76510 | 2.5615 | 3.5990 | 4.76875 | 8.07962 |
| **max_spent_in_single_shopping** | 210.0 | 5.40807 | 0.4914 | 4.51900 | 5.0450 | 5.2230 | 5.87700 | 6.55000 |

Table No. 4

***As observed in the above descriptive statistics, the range of fields vary from one another. Hence, we would require scaling for clustering.***
After applying zscore method of scaling, the scaled data detail is as follows:

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **spending** | 210.0 | 9.148766E- | 1.00238 | -1.4667 | -0.8879 | -0.1696 | 0.8465 | 2.1815 |
| **advance_payments** | 210.0 | 1.097006E- | 1.00238 | -1.6496 | -0.8514 | -0.1836 | 0.8870 | 2.0652 |
| **probability_of_full_payment** | 210.0 | 1.638372E- | 1.00238 | -2.5713 | -0.6009 | 0.10317 | 0.7126 | 2.0113 |
| **current_balance** | 210.0 | -1.358702E | 1.00238 | -1.6505 | -0.8286 | -0.2376 | 0.7945 | 2.3675 |
| **credit_limit** | 210.0 | -2.790757E | 1.00238 | -1.6682 | -0.8349 | -0.0573 | 0.8044 | 2.0551 |
| **min_payment_amt** | 210.0 | 1.554312E- | 1.00238 | -1.9664 | -0.7616 | -0.0659 | 0.7185 | 2.9389 |
| **max_spent_in_single_shopping** | 210.0 | -1.935489E | 1.00238 | -1.8132 | -0.7404 | -0.3774 | 0.9563 | 2.3289 |

Table No. 5

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

We can apply hierarchical clustering to scaled data. Then we had plotted dendrogram to find the optimum numbers of clusters.

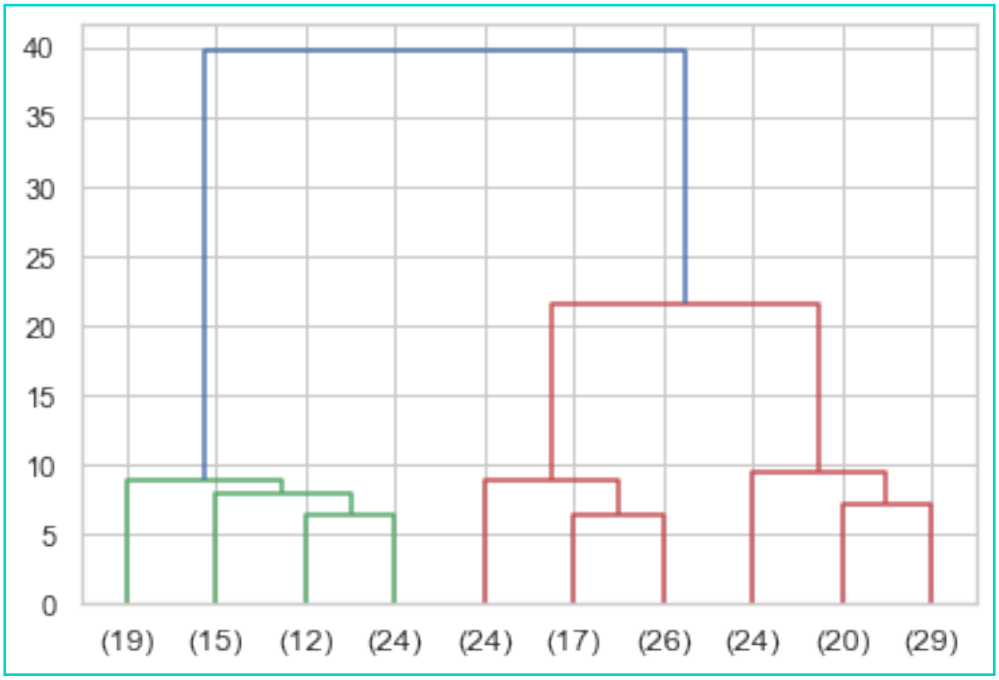



Figure No. 6

From the above plot, we can divide the dataframe into 3 clusters as those 3 are the big clusters:

| clusters | spending | advance_payments | probability_of_full_payment | current_balance | credit_limit | min_payment_amt | max_spent_in_single_shopping |
|---|---|---|---|---|---|---|---|
| **1** | 1286.00 | 1130.18 | 61.908000 | 431.072 | 257.924 | 254.74100 | 421.216 |
| **2** | 795.45 | 888.22 | 56.826363 | 351.009 | 190.852 | 331.00025 | 343.188 |
| **3** | 1036.53 | 1039.05 | 64.180900 | 399.911 | 235.531 | 190.68920 | 371.291 |

Table No. 6

**Observations:¶**

1. Cluster 1
    - Highest : spending, advance_payments, current_balance, credit_limit and max_spent_in_single_shopping
    - Moderate : probability_of_full_payment, min_payment_amt.
2. Cluster 2
    - Highest : min_payment_amt.
    - Lowest : spending, advance_payments, current_balance, credit_limit, max_spent_in_single_shopping and probability_of_full_payment
3. Cluster 3
    - Highest : probability_of_full_payment.
    - Moderate : spending, advance_payments, current_balance, credit_limit and max_spent_in_single_shopping.
    - Lowest : min_payment_amt

Graphical representation of the same is as follows:

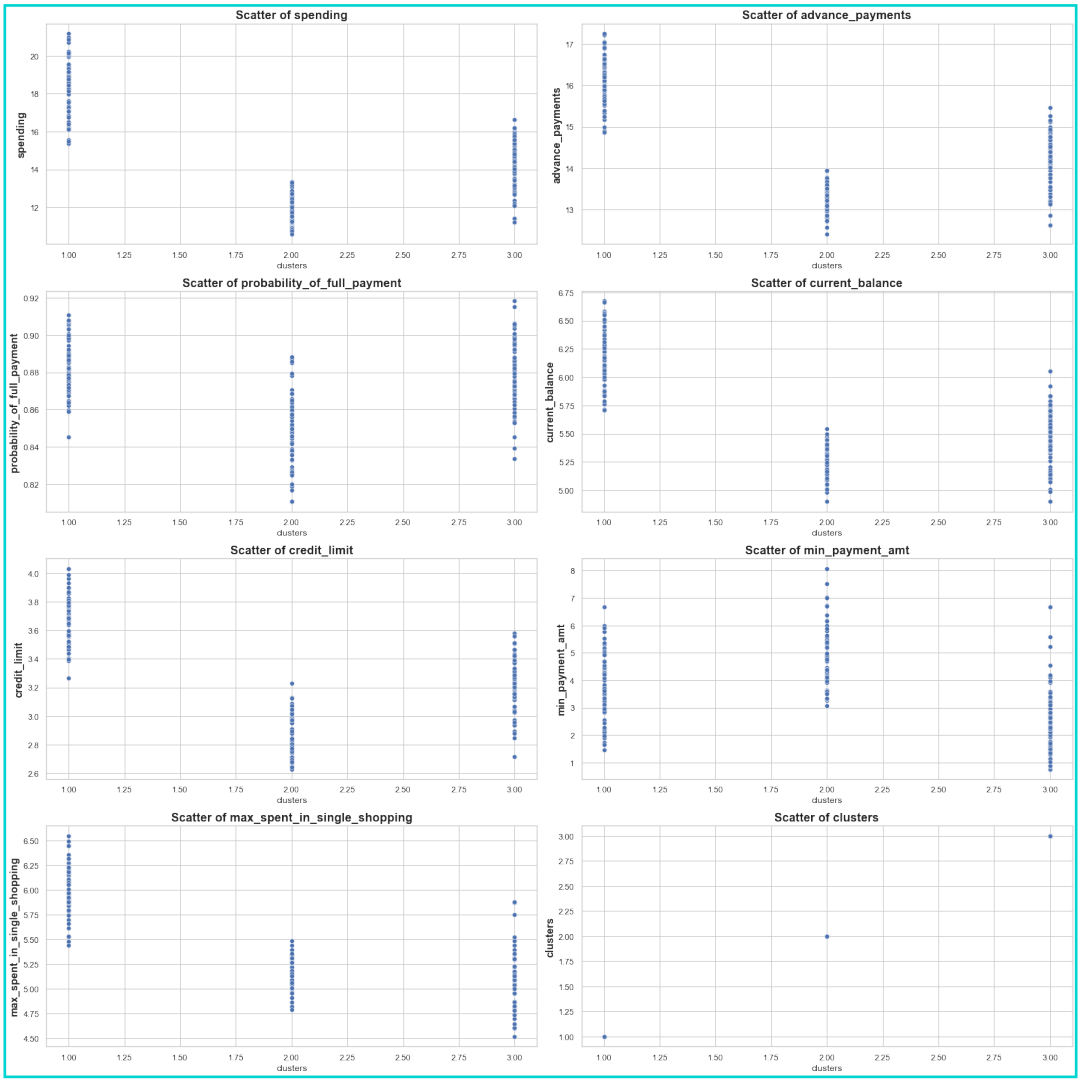



Figure No. 7

## IV. 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 finalised clusters.

We can apply K-Means clustering to scaled data. First of all, Then we can find KMeans.inertia_ and silhouette_score for various no. of clusters along with the elbow graph.

| Inertia | silhouette_score | Number of Clusters |
|---|---|---|
| 708.499737 | 0.496583 | 2 |
| 441.127253 | 0.451895 | 3 |
| 390.103578 | 0.363339 | 4 |
| 343.154651 | 0.306584 | 5 |
| 308.725421 | 0.281221 | 6 |
| 276.122717 | 0.275013 | 7 |
| 254.475792 | 0.260384 | 8 |
| 234.102560 | 0.265238 | 9 |
| 213.966970 | 0.262797 | 10 |

Table No. 7

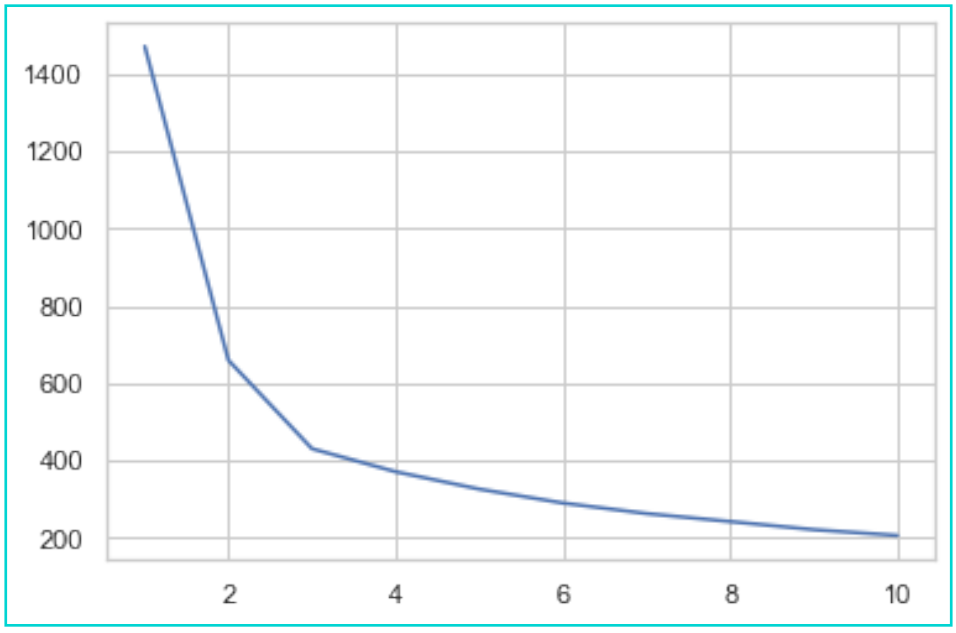



Figure No. 8

From the above graph, I had tested for both no. of clusters = 2 and 3. Since the silhouette_score i.e. (0.4965830066767461) is better for no. of clusters = 2, hence I chose it

## Observations:¶

1. Cluster 0
   All the values in this cluster is at lower end.
2. Cluster 1
   All the values in this cluster is at higher end.

| K_Clusters | spending | advance_payments | probability_of_full_payment | current_balance | credit_limit | min_payment_amt | max_spent_in_single_shopping |
|---|---|---|---|---|---|---|---|
| **0** | 1302.63 | 1145.64 | 62.782700 | 437.125 | 261.389 | 256.78100 | 427.093 |
| **1** | 1815.35 | 1911.81 | 120.13256 | 744.867 | 422.918 | 519.64945 | 708.602 |

Table No. 8

Graphical representation of the same is as follows:

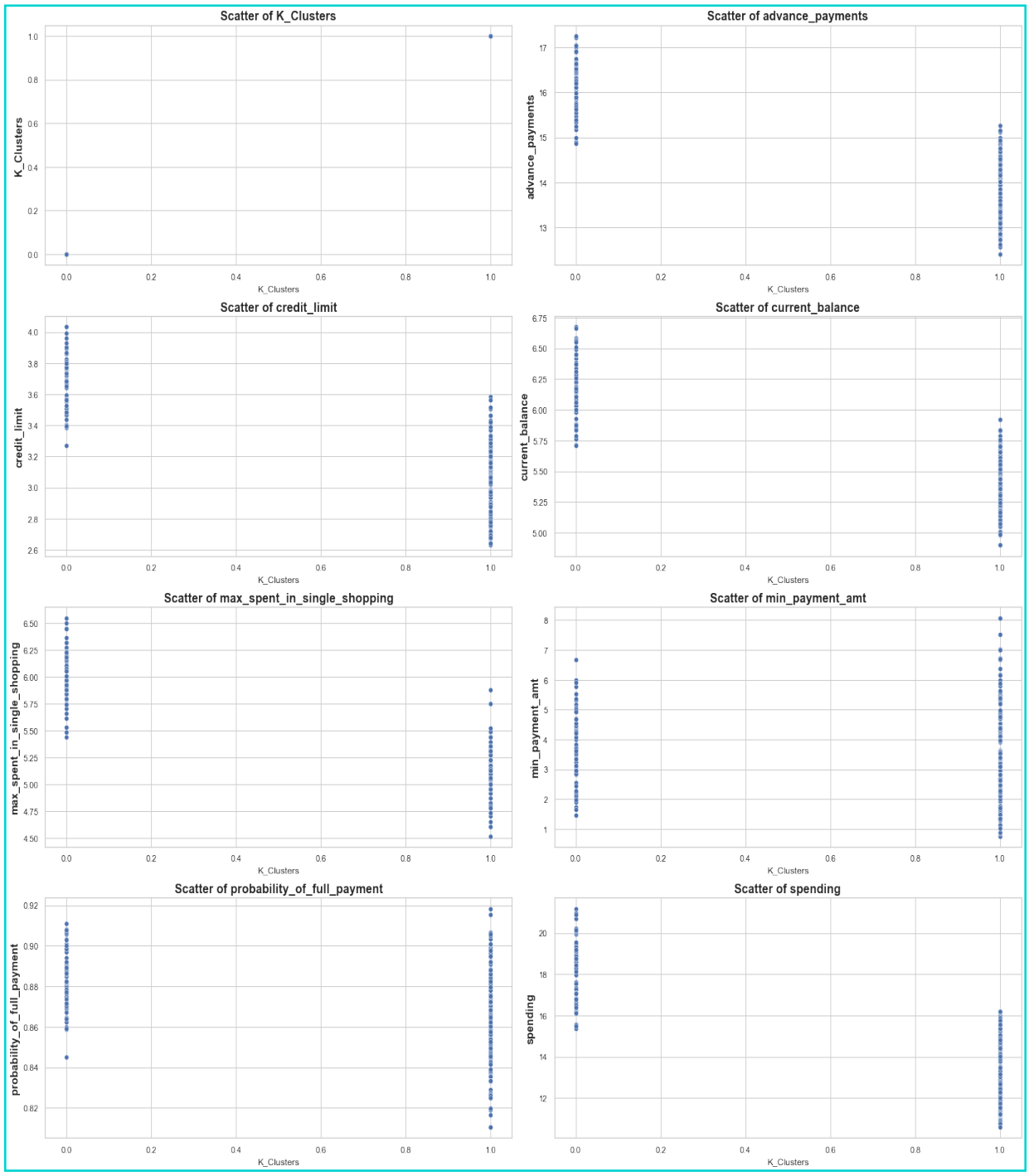

Scatter of K_Clusters
K_Clusters
Scatter of advance_payments
advance_payments
Scatter of credit_limit
credit_limit
Scatter of current_balance
current_balance
Scatter of max_spent_in_single_shopping
max_spent_in_single_shopping
Scatter of min_payment_amt
min_payment_amt
Scatter of probability_of_full_payment
probability_of_full_payment
Scatter of spending
spending

Figure No. 9

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

To describe the cluster profiles for 2 clusters, we will look into the clustered information in details:

**Spending & Advance_payments**

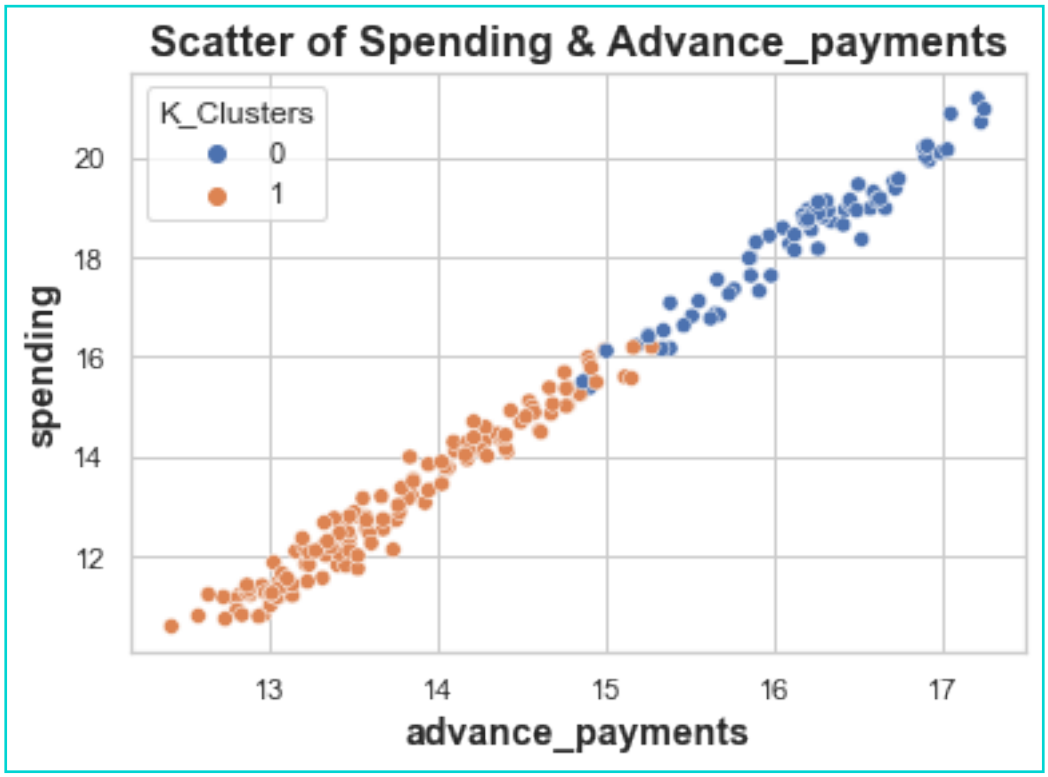



Figure No. 10.1

| K_Clusters | spending | advance_payments |
|---|---|---|
| 0 | 18.346901 | 16.135775 |
| 1 | 13.060072 | 13.754029 |

Table No. 9. 1

## Observations

- Cluster 0 : It divides the dataframe into high spending and high advance_payments.
- Cluster 1 : It divides the dataframe into low spending and low advance_payments.
- The clusters are not well separated.

## Advance_payments & probability_of_full_payment

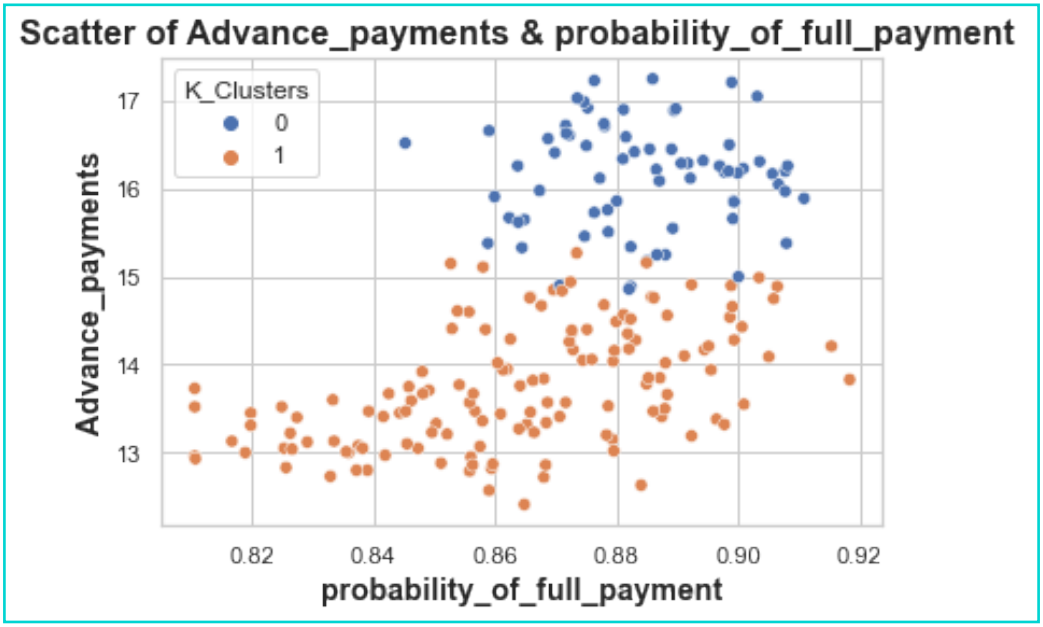



Figure No. 10. 2

| K_Clusters | advance_payments | probability_of_full_payment |
|---|---|---|
| 0 | 16.135775 | 0.884263 |
| 1 | 13.754029 | 0.864263 |

Table No. 9. 2

## Observations

- Cluster 0 : It divides the dataframe into high probability_of_full_payment and high advance_payments.
- Cluster 1 : It divides the dataframe into low probability_of_full_payment and low advance_payments.
- The clusters are not well separated.
- The clusters overlap.

## Current_balance & probability_of_full_payment

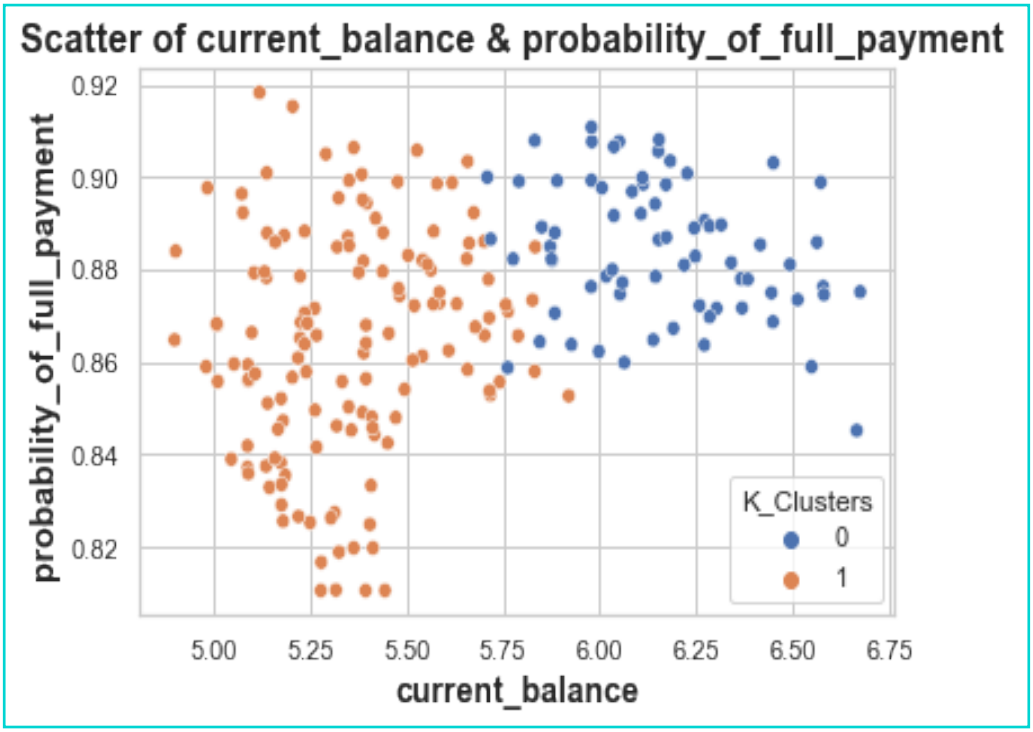



Figure No. 10.3

| K_Clusters | probability_of_full_payment | current_balance |
|---|---|---|
| 0 | 0.884263 | 6.156690 |
| 1 | 0.864263 | 5.358755 |

Table No. 9.3

## Observations

- Cluster 0 : It divides the dataframe into high probability_of_full_payment and high Current_balance
- Cluster 1 : It divides the dataframe into low probability_of_full_payment and low Current_balance.
- The clusters are not well separated.
- The clusters overlap.

## Current_balance & Credit_limit

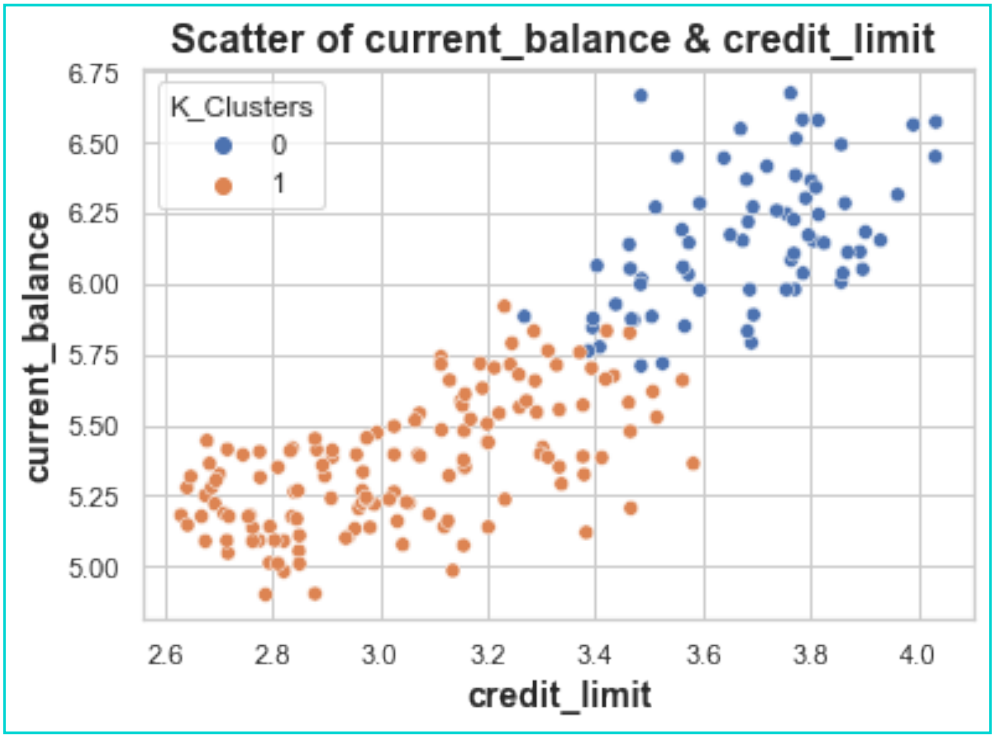



Figure No. 10.4

| K_Clusters | current_balance | credit_limit |
|---|---|---|
| 0 | 6.156690 | 3.681535 |
| 1 | 5.358755 | 3.042576 |

Table No. 9.4

## Observations

- Cluster 0 : It divides the dataframe into low Current_balance & Credit_limit.
- Cluster 1 : It divides the dataframe into high Current_balance & Credit_limit.
- The clusters are not well separated.
- The clusters overlap.

## Credit_limit & Min_payment_amt

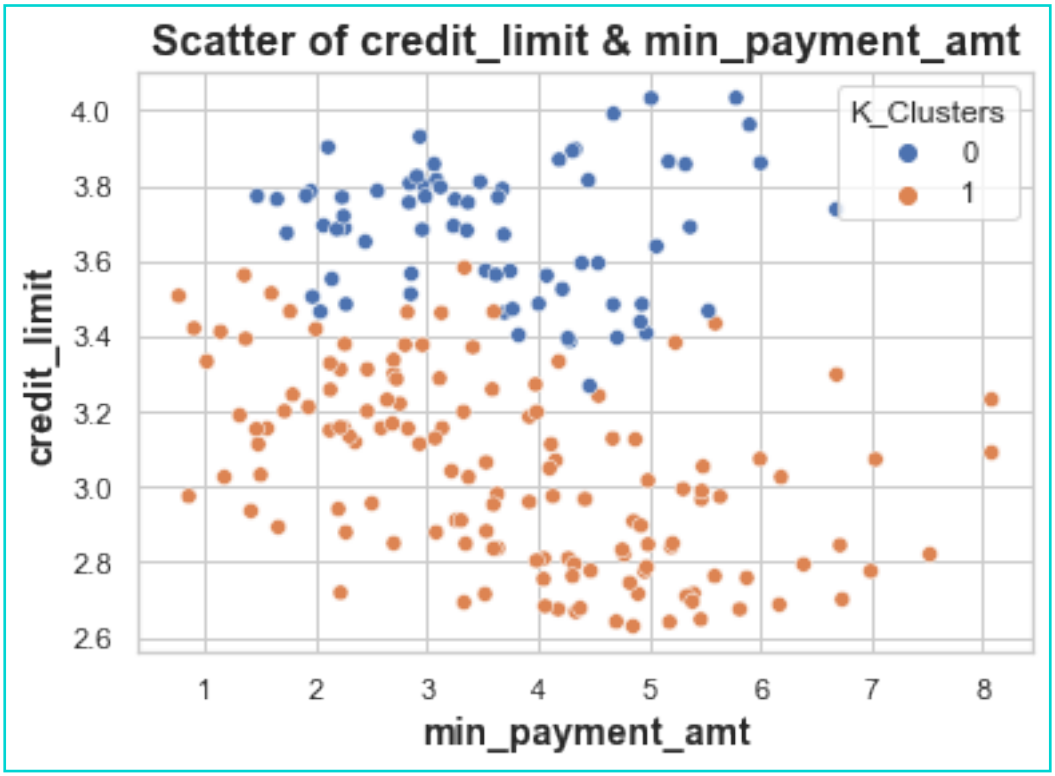



Figure No. 10.5

| K_Clusters | credit_limit | min_payment_amt |
|---|---|---|
| 0 | 3.681535 | 3.616634 |
| 1 | 3.042576 | 3.738485 |

Table No. 9.5

## Observations

- Cluster 0 : It divides the dataframe into low Credit_limit & Min_payment_amt .
- Cluster 1 : It divides the dataframe into high Credit_limit & Min_payment_amt .
- The clusters are not well separated.
- The clusters overlap.

## Min_payment_amt & max_spent_in_single_shopping

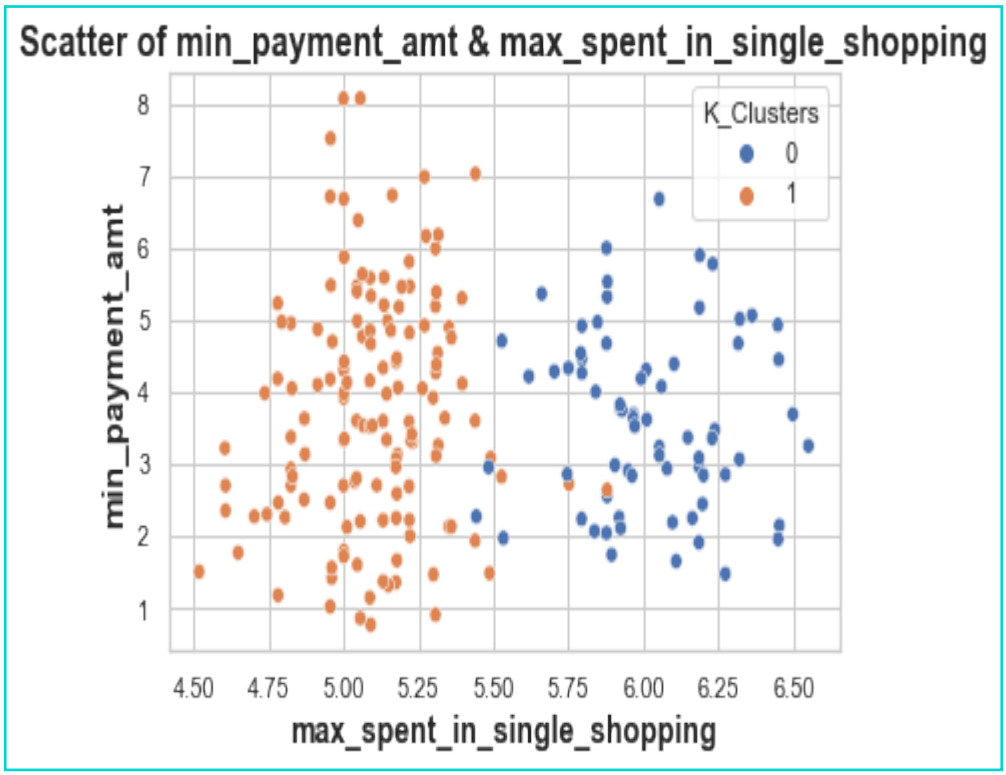



Figure No. 10.6

| K_Clusters | min_payment_amt | max_spent_in_single_shopping |
|---|---|---|
| 0 | 3.616634 | 6.015394 |
| 1 | 3.738485 | 5.097856 |

Table No. 9.6

## Observations

- Cluster 0 : It divides the dataframe into low Min_payment_amt & max_spent_in_single_shopping.
- Cluster 1 : It divides the dataframe into high Min_payment_amt & max_spent_in_single_shopping.
- The clusters are not well separated.
- The clusters overlap.

## Recommendations

1. The customers with cluster 0, is a high spending group, so more products with higher price can be promoted to this group.
2. Credit limit can be increased for the customer group with cluster 0, as they are higher spending group as well as the advance payments made by this group is also higher. Hence, if we increase the credit limit more , the spending might increase further.
3. The probability that the amount will be paid is almost same for both the groups and is pretty good. Hence more credit cards can be offered like :
    1. For Group 0( cluster 0) : Higher limit credit cards are be promoted.
    2. For Group 1 ( cluster 1) : Lower limit credit cards can be promoted.
4. Personal Loans can be promoted to group 1, as they make more minimum payments than group 0, which can help them repay the credit cards.

# PROBLEM - 2 SUMMARY

## INSURANCE FIRM ANALYSIS

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.

### Exploratory Data Analysis

#### Data Description

#### Data Dictionary for Insurance Firm Analysis

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)

# PROBLEM 2

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

### Sample of the dataset

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

Table No. 10.1

### Data types of different variable

| Column | Dtype |
|---|---|
| Age | int64 |
| Agency_Code | object |
| Type | object |
| Claimed | object |
| Commision | float64 |
| Channel | object |
| Duration | int64 |
| Sales | float64 |
| Product Name | object |
| Destination | object |

Table No. 10.2

## Observations¶

1. Dataset has a total of 3000 rows and 10 columns.
2. There are 2 float64 , 2 int64 and 6 object types.
3. There is no missing values.

## Checking for duplicates

There is 139 duplicates present in the dataframe.
**Example duplicate rows :**

| | Age | Agency_Code | Type | Claimed | Commission | Channel | Duration | Sales | Product Name | Destination |
|---|---|---|---|---|---|---|---|---|---|---|
| **63** | 30 | C2B | Airlines | Yes | 15.0 | Online | 27 | 60.0 | Bronze Plan | ASIA |
| **329** | 36 | EPX | Travel Agency | No | 0.0 | Online | 5 | 20.0 | Customised Plan | ASIA |
| **407** | 36 | EPX | Travel Agency | No | 0.0 | Online | 11 | 19.0 | Cancellation Plan | ASIA |
| **411** | 35 | EPX | Travel Agency | No | 0.0 | Online | 2 | 20.0 | Customised Plan | ASIA |
| **422** | 36 | EPX | Travel Agency | No | 0.0 | Online | 5 | 20.0 | Customised Plan | ASIA |

Table No. 10.3

Hence all the duplicate rows have been dropped.

## Missing data analysis for the variables

No missing values present in the data frame.

## Descriptive Statistics

| Fields | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Age** | 2861.0 | 38.204124 | 10.678106 | 8.0 | 31.0 | 36.00 | 43.00 | 84.00 |
| **Commision** | 2861.0 | 15.080996 | 25.826834 | 0.0 | 0.0 | 5.63 | 17.82 | 210.21 |
| **Duration** | 2861.0 | 72.120238 | 135.977200 | -1.0 | 12.0 | 28.00 | 66.00 | 4580.00 |
| **Sales** | 2861.0 | 61.757878 | 71.399740 | 0.0 | 20.0 | 33.50 | 69.30 | 539.00 |

Table No. 10.4

## Observation

1. Total No. of column after removing duplicate columns is 2861.
2. Means have slightly different range, hence scaling will be useful.
3. Duration should be at least 1 day, hence -1 is bad data and seems to have outlier too.
4. Sales and Commission also seems to have outliers too.

After removal of bad data, the descriptive statistics is as follows :

| Fields | count | mean | std | min | 25% | 50% | 75% | max | Sales |
|---|---|---|---|---|---|---|---|---|---|
| **Age** | 2858.0 | 38.205738 | 10.679258 | 8.0 | 31.0 | 36.00 | 43.00 | 84.00 | 20.0 |
| **Commision** | 2858.0 | 15.077218 | 25.830284 | 0.0 | 0.0 | 5.63 | 17.82 | 210.21 | 19.0 |
| **Duration** | 2858.0 | 72.196291 | 136.028290 | 1.0 | 12.0 | 28.00 | 66.00 | 4580.00 | 20.0 |
| **Sales** | 2858.0 | 61.772841 | 71.413840 | 0.0 | 20.0 | 33.50 | 69.30 | 539.00 | 20.0 |

Table No. 10.5

## Univariate Analysis

### Histogram for Numerical Fields

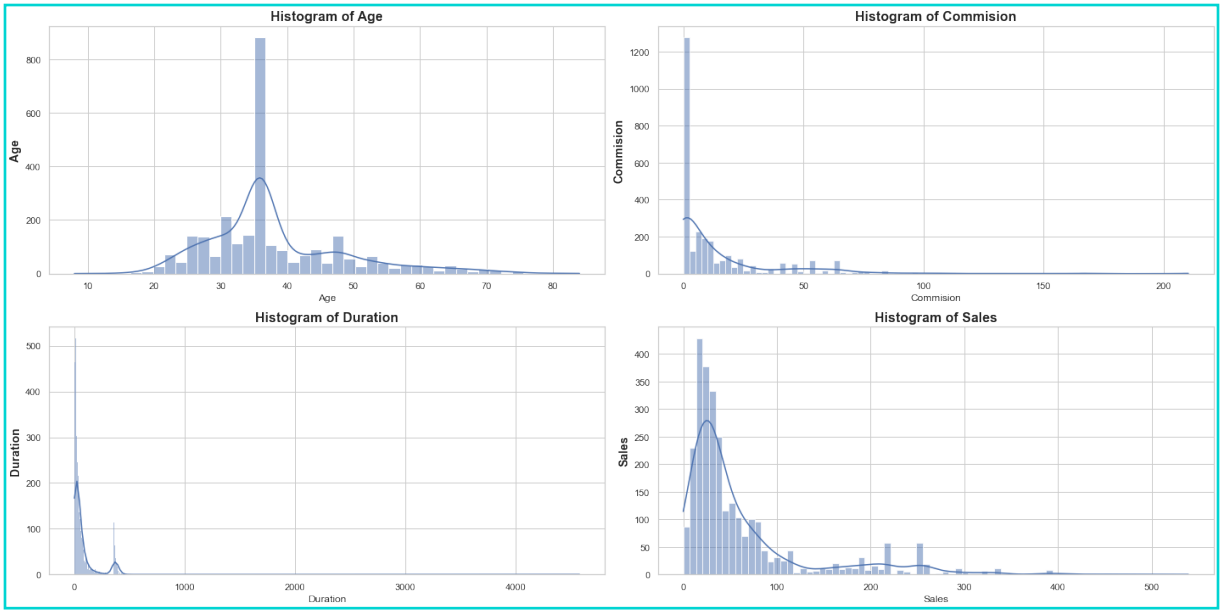



Figure No. 11.1

## Observations

1. Outliers can be observed in all the plots.
2. Skewness :
   - **A. Positively Skewed**
     - Skewness of Commission is 3.1031126292410716.
     - Skewness of Duration is 13.778867077621834
     - Skewness of Sales is 2.3434132352067008
   - **B. Approximate Normal Distribution**
     - Skewness of Age is 1.1025661500650201

## CountPlot of Categorical Fields

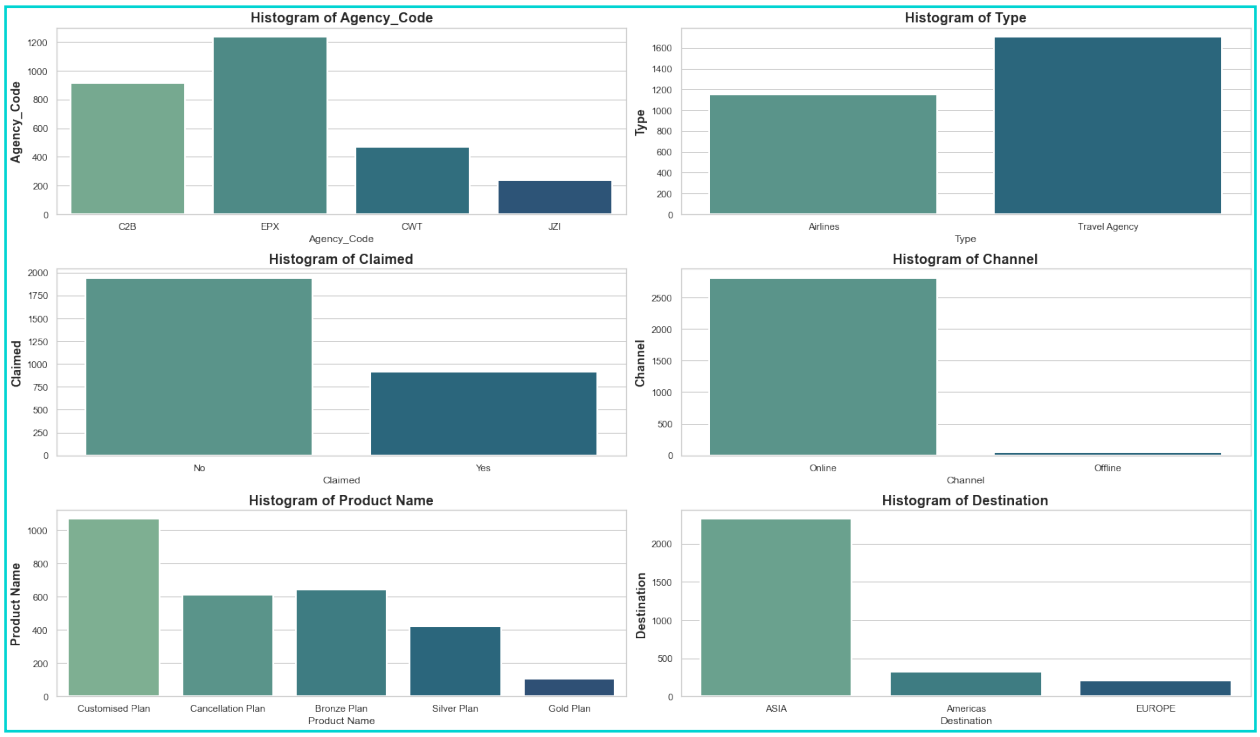



Figure No. 11.2

## Observations

1. In Agency_Code, highest count is off EPX and lowest is of JZ1.
2. We have maximum counts for Travel Agency among Airlines & Travel Agency.
3. Most number of Insurance is not claimed.
4. Most preferred channel is Online.
5. Customised Plan has been bought the most.
6. Maximum number of destination is Asia.

## Bivariant Analysis¶

This can be done both by Correlation Heat Map and PairPlot

### HeatMap

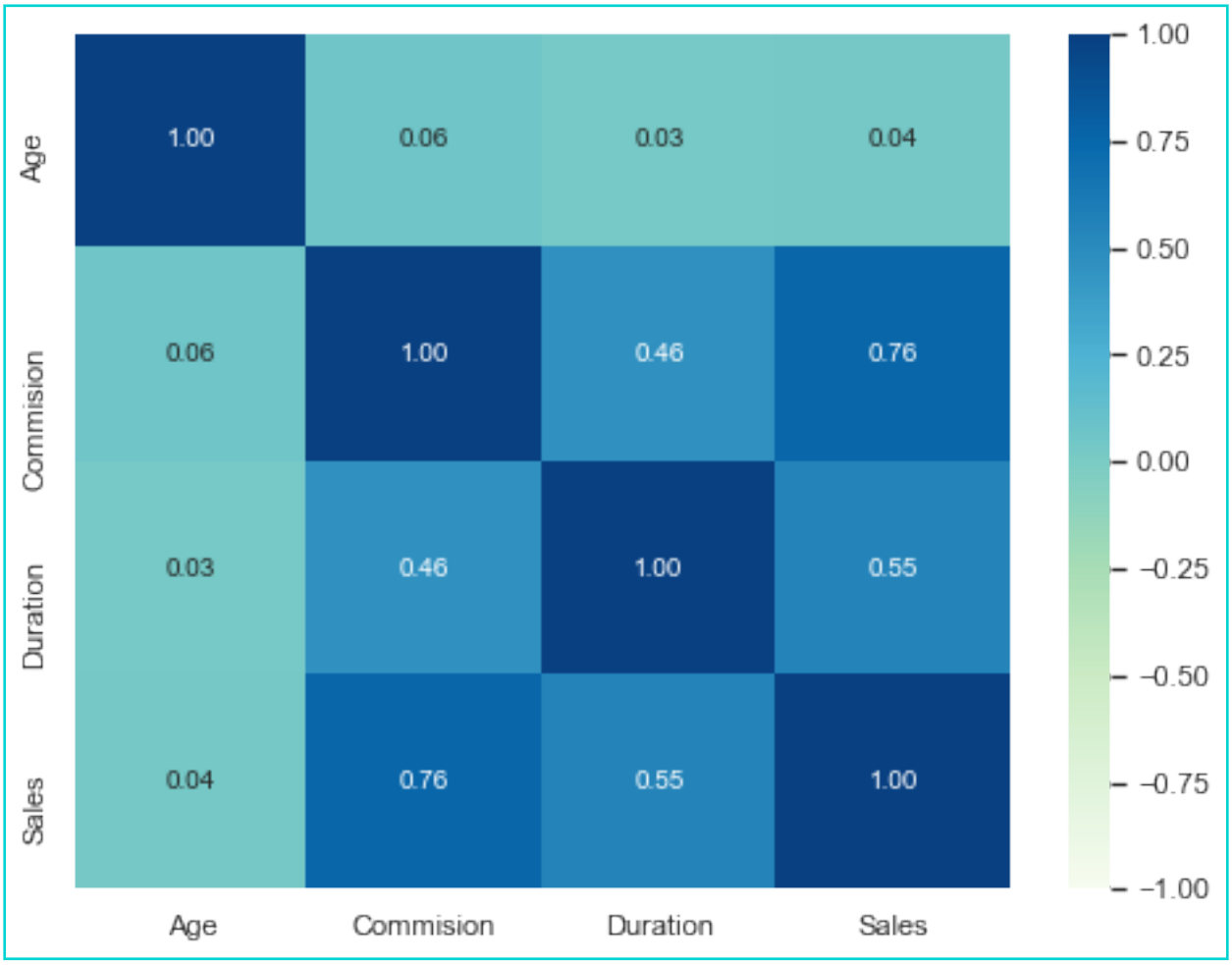



Figure No. 11.3

## Observations¶

1. Some correlation can be observed between commision-Sales.
2. Other then that all other correlations are not significant.

## PairPlot

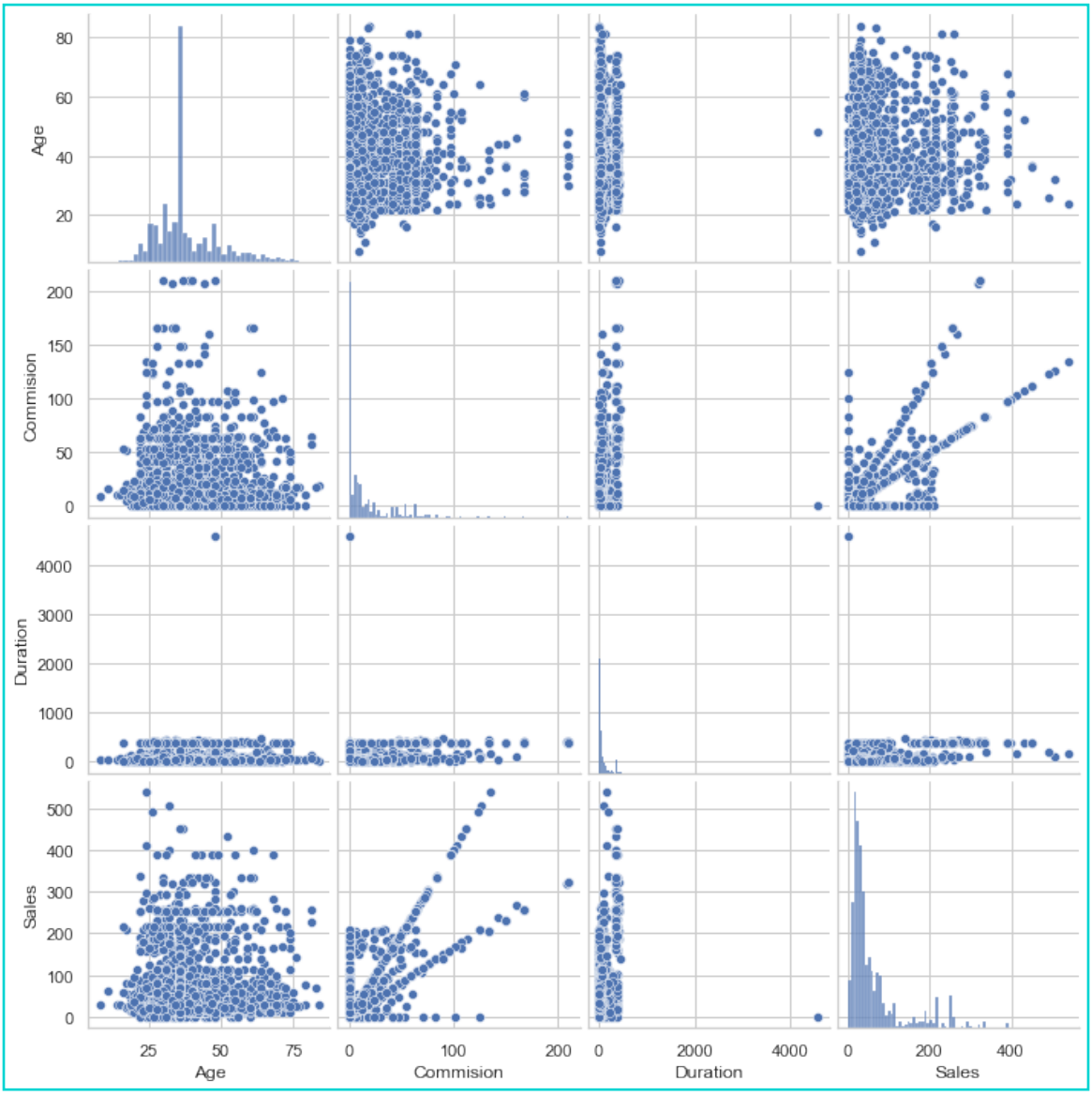



Figure No. 11.4

## Observations:¶

1. Positive covariance can be observed in commision-Sales.
2. Negative Covariance is not observed.

# BoxPlot

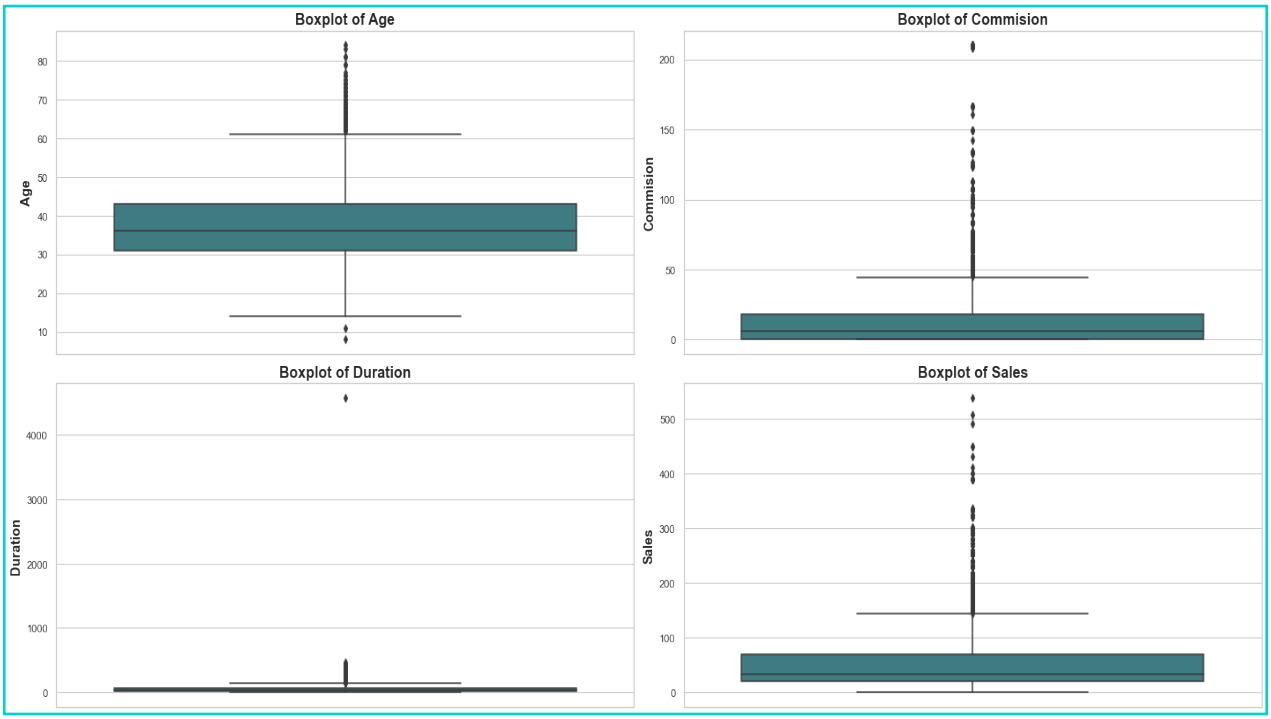



Figure No. 11.5

## Observations:¶

1. Outliers can be observed in Age, Commission, Duration and Sales.
2. Outlier treatment can be done for the mentioned fields.

After treating Outliers, we observe the below Box plots.

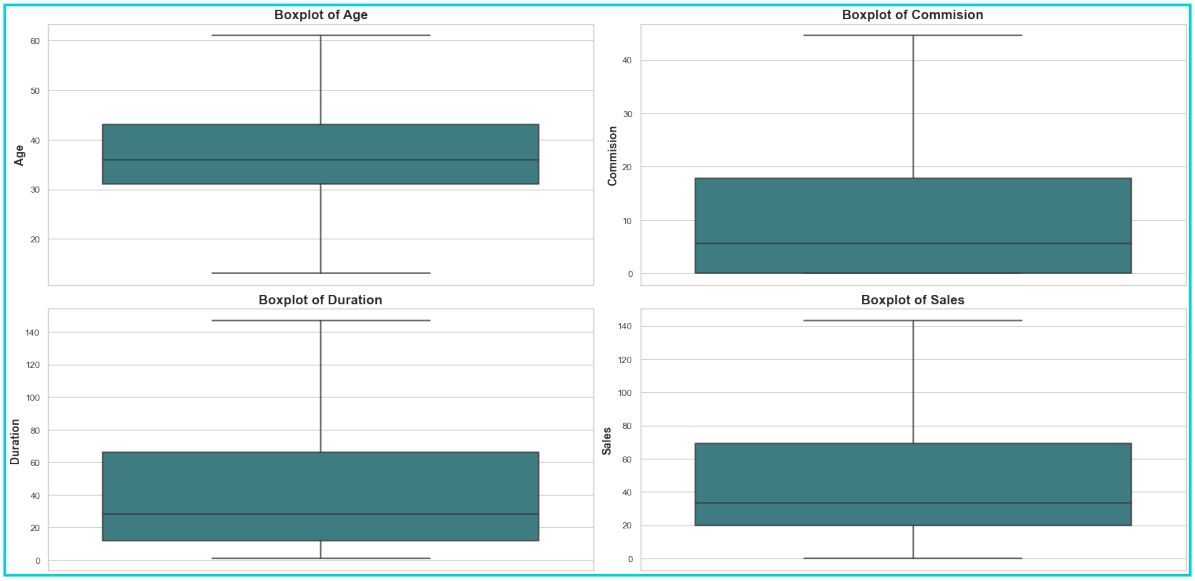



Figure No. 11.6

## Scaling

As observed in the descriptive statistics, we can say that scalling should be done.

| Fields | count | mean | std | min | 25% | 50% | 75% | max | Sales |
|---|---|---|---|---|---|---|---|---|---|
| Age | 2858.0 | 38.205738 | 10.679258 | 8.0 | 31.0 | 36.00 | 43.00 | 84.00 | 20.0 |
| Commision | 2858.0 | 15.077218 | 25.830284 | 0.0 | 0.0 | 5.63 | 17.82 | 210.21 | 19.0 |
| Duration | 2858.0 | 72.196291 | 136.028290 | 1.0 | 12.0 | 28.00 | 66.00 | 4580.00 | 20.0 |
| Sales | 2858.0 | 61.772841 | 71.413840 | 0.0 | 20.0 | 33.50 | 69.30 | 539.00 | 20.0 |

Table No. 10.6

Hence, after applying the z-score scaling, the statistics is as follows:

| Fields | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Age | 2858.0 | 1.818388E-16 | 1.000175 | -2.535324 | -0.702374 | -0.193222 | 0.519592 | 2.352543 |
| Commision | 2858.0 | 2.340541E-17 | 1.000175 | -0.758443 | -0.758443 | -0.395078 | 0.391675 | 2.116852 |
| Duration | 2858.0 | 2.159846E-17 | 1.000175 | -0.980968 | -0.748357 | -0.410012 | 0.393556 | 2.106426 |
| Sales | 2858.0 | 1.015633E-16 | 1.000175 | -1.199765 | -0.730087 | -0.413055 | 0.427669 | 2.164303 |

Table No. 10.7

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

In our dataframe, we have a mixture of Data Types and object data should be converted into categorical/numerical data to fit in the models. Hence all the object type fields i.e. Agency_Code, Type, Claimed, Channel, Product Name and Destination are converted to codes.

The new datatypes is as follows:

| Column | Dtype |
|---|---|
| Agency_Code | int8 |
| Type | int8 |
| Claimed | int8 |
| Commision | float64 |
| Channel | int8 |
| Duration | float64 |
| Sales | float64 |
| Product Name | int8 |
| Destination | int8 |

Table No. 11.1

### Observations

1. There are 2858 rows and 10 rows.
2. There are float64(4), int8(6).

### Proportion of 1s and 0s

| index | Claimed |
|---|---|
| 0 | 0.680196 |
| 1 | 0.319804 |

Figure No. 11.2

The sample data looks as :

| Age | Agency_Code | Type | Claimed | Commission | Channel | Duration | Sales | Product Name | Destination |
|---|---|---|---|---|---|---|---|---|---|
| 48.0 | 0 | 0 | 0 | 0.70 | 1 | 7.0 | 2.51 | 2 | 0 |
| 36.0 | 2 | 1 | 0 | 0.00 | 1 | 34.0 | 20.00 | 2 | 0 |
| 39.0 | 1 | 1 | 0 | 5.94 | 1 | 3.0 | 9.90 | 2 | 1 |
| 36.0 | 2 | 1 | 0 | 0.00 | 1 | 4.0 | 26.00 | 1 | 0 |
| 33.0 | 3 | 0 | 0 | 6.30 | 1 | 53.0 | 18.00 | 0 | 0 |

Table No. 11.3

Extracting the target column into separate vectors for training set and test set, where X is data without output column and y with only target/output column
Then after splitting data into Train and Test, with 70% as Train and 30% as Test, I have the following shape of the created dataframe. I have splitted the train data into 70% as it will be adequate amount to train the model where 30% data is good enough for testing the model.

| Data Frame | Number of Rows | Number of Columns |
|---|---|---|
| X_train | 2000 | 9 |
| X_test | 858 | 9 |
| train_labels | 2000 | 1 |
| test_labels | 858 | 1 |

Table No. 11.4

I have created 3 models: Cart, Random Forest and ANN. The details of each model is as follows:

## CART

In this model, we had appleid GridSearchCV, to find the best parameter and is as follows:

### Parameters

1. Criterion : gini
2. Max_depth : 10
3. Min_samples_leaf : 20
4. Min_samples_split : 180

### Variable Importance

| Variables | Imp |
|---|---|
| Agency_Code | 0.544634 |
| Sales | 0.195672 |
| Product Name | 0.120670 |
| Duration | 0.067801 |
| Commision | 0.033230 |
| Age | 0.030816 |
| Destination | 0.007178 |
| Type | 0.000000 |
| Channel | 0.000000 |

Table No. 11.5

**Agency_Code is the most important variable for predicting claims.**

## RANDOM Forest

Grid Search for finding out the optimal values for the hyper parameters¶
Due to large volume of data, trying for different parameter values in the grid search with higher cv value will lead to performance issues and model will run for much longer time.

### Parameters

1. n_estimators : 300
2. Max_depth : 10
3. Min_samples_leaf : 10
4. Min_samples_split : 50
5. Max_features : 6

## Variable Importance

| Variable | Imp |
|---|---|
| Agency_Code | 0.351272 |
| Sales | 0.196934 |
| Product Name | 0.171991 |
| Duration | 0.100257 |
| Commision | 0.080783 |
| Age | 0.069449 |
| Type | 0.015376 |
| Destination | 0.012187 |
| Channel | 0.001751 |

Table No. 11.6

**Agency_Code is the most important variable for predicting claims.**

## Artificial Neural Network

Grid Search for finding out the optimal values for the hyper parameters¶
Due to large volume of data, trying for different parameter values in the grid search with higher cv value will lead to performance issues and model will run for much longer time.

### Parameters

1. hidden_layer_sizes : 200
2. max_iter : 7000
3. solver : adam
4. tol : 0.01

## III. 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.

### Model Evaluation of CART

### Training data

#### ROC Curve

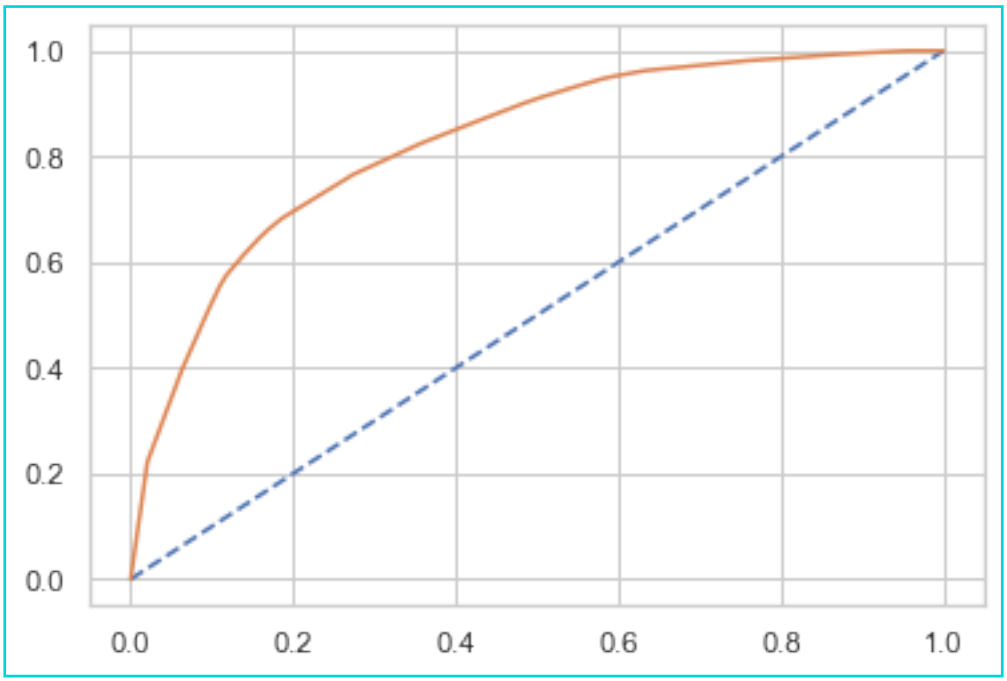



Figure No. 12.1

#### Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 1200 | 160 |
| | Yes | 272 | 368 |

Table No. 12. 1

## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.88 | 0.85 | 1360 |
| 1 | 0.70 | 0.57 | 0.63 | 640 |
| accuracy | | | 0.78 | 2000 |
| macro avg | 0.76 | 0.73 | 0.74 | 2000 |
| weighted avg | 0.78 | 0.78 | 0.78 | 2000 |

Table No. 12.2

# Test data

## ROC Curve

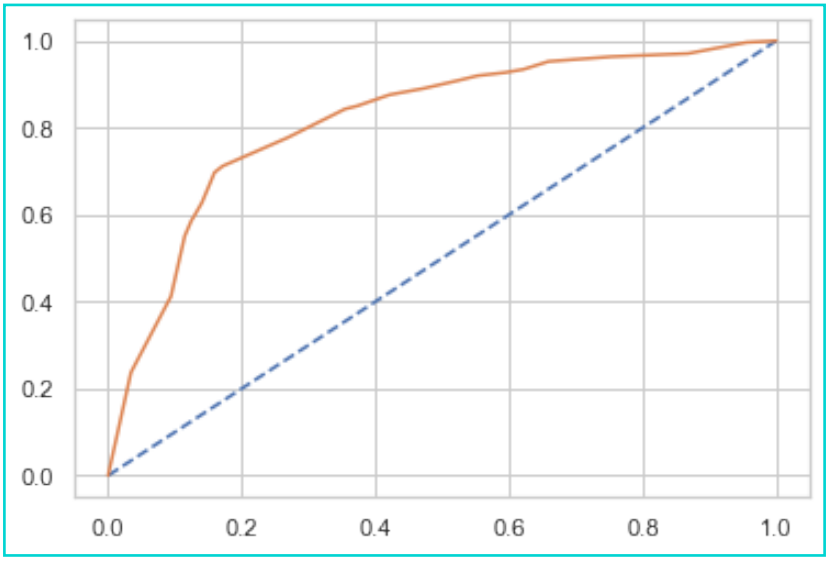



Figure No. 12.2

## Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 510 | 74 |
| | Yes | 112 | 162 |

Table No. 12. 3

## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.87 | 0.85 | 584 |
| 1 | 0.69 | 0.59 | 0.64 | 274 |
| accuracy | | | 0.78 | 858 |
| macro avg | 0.75 | 0.73 | 0.74 | 858 |
| weighted avg | 0.78 | 0.78 | 0.78 | 858 |

Table No. 12.4

## Cart Conclusion

Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

| | Train Data | Test Data |
|---|---|---|
| AUC | 82.6% | 81.8% |
| Accuracy | 78.4% | 78.3% |
| Sensitivity | 88.2% | 87.3% |
| Precision | 70% | 69% |
| Recall | 57% | 59% |
| f1-Score | 63% | 64% |

Table No. 12.5

## Model Evaluation of Random Forest

## Training data

### Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 1217 | 143 |
| | Yes | 261 | 379 |

Table No. 12. 6

## ROC Curve

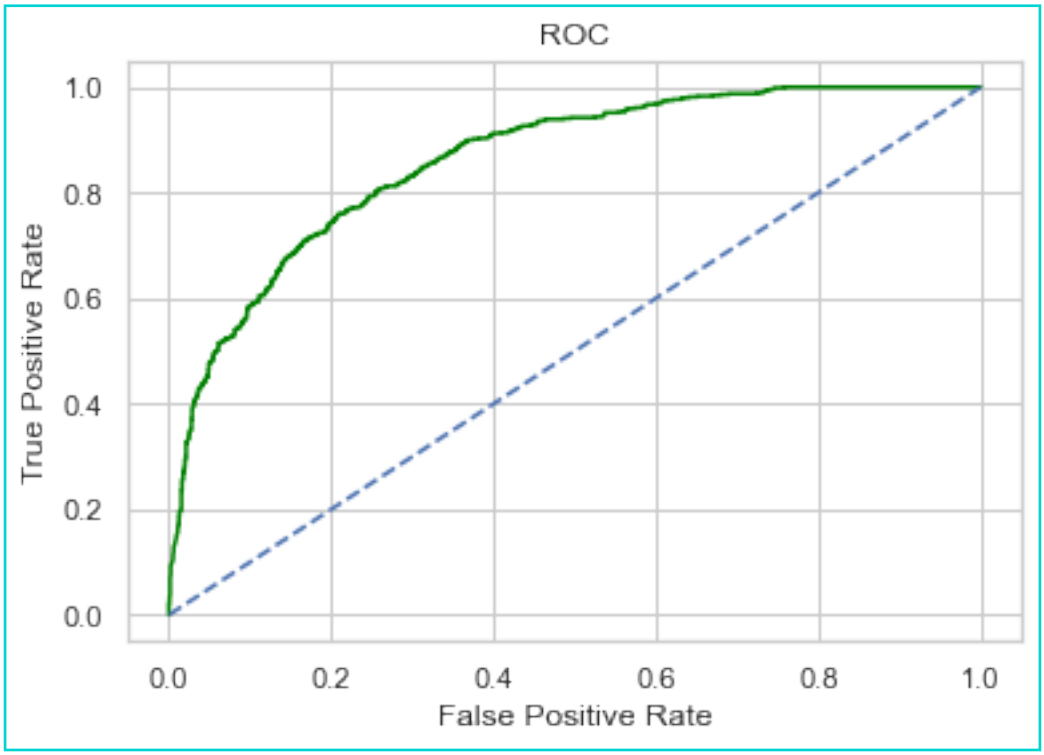



Figure No. 12.3

## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.82 | 0.89 | 0.86 | 1360 |
| 1 | 0.73 | 0.59 | 0.65 | 640 |
| accuracy | | | 0.80 | 2000 |
| macro avg | 0.77 | 0.74 | 0.75 | 2000 |
| weighted avg | 0.79 | 0.80 | 0.79 | 2000 |

Table No. 12.7

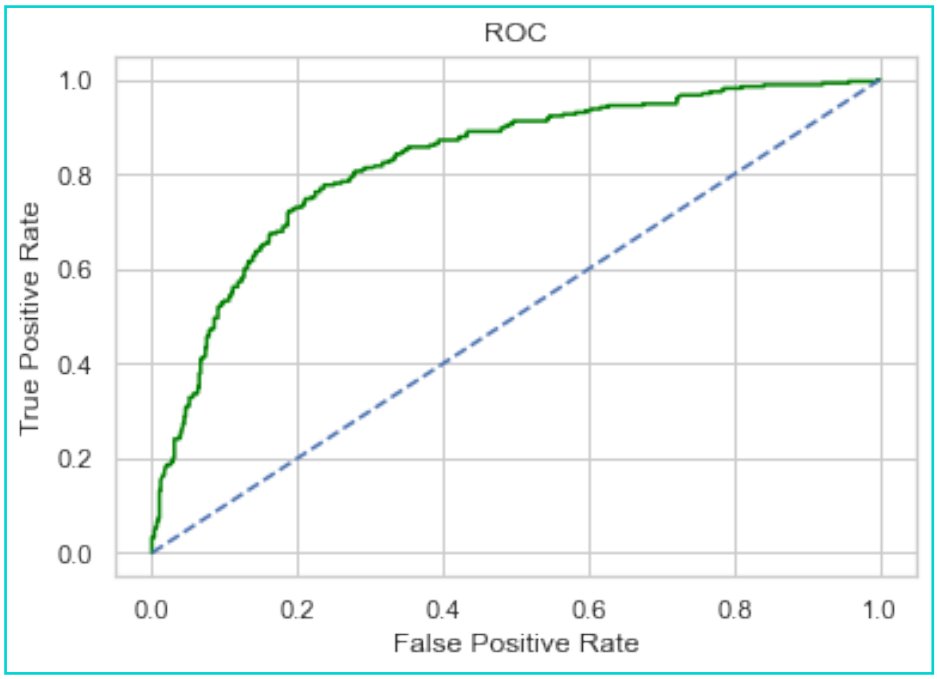
# Testing data

## Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 506 | 78 |
| | Yes | 105 | 169 |

Table No. 12. 8

## ROC Curve




# Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.87 | 0.85 | 584 |
| 1 | 0.68 | 0.62 | 0.65 | 274 |
| accuracy | | | 0.79 | 858 |
| macro avg | 0.76 | 0.74 | 0.75 | 858 |
| weighted avg | 0.78 | 0.79 | 0.78 | 858 |

Table No. 12.9

## RM Conclusion

Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

| | Train Data | Test Data |
|---|---|---|
| AUC | 74.3% | 74.1% |
| Accuracy | 79.8% | 78.6% |
| Sensitivity | 89.4% | 86.6% |
| Precision | 73% | 68% |
| Recall | 59% | 62% |
| f1-Score | 65% | 65% |

Table No. 12.10

## Model Evaluation of ANN

## Training data

### ROC Curve

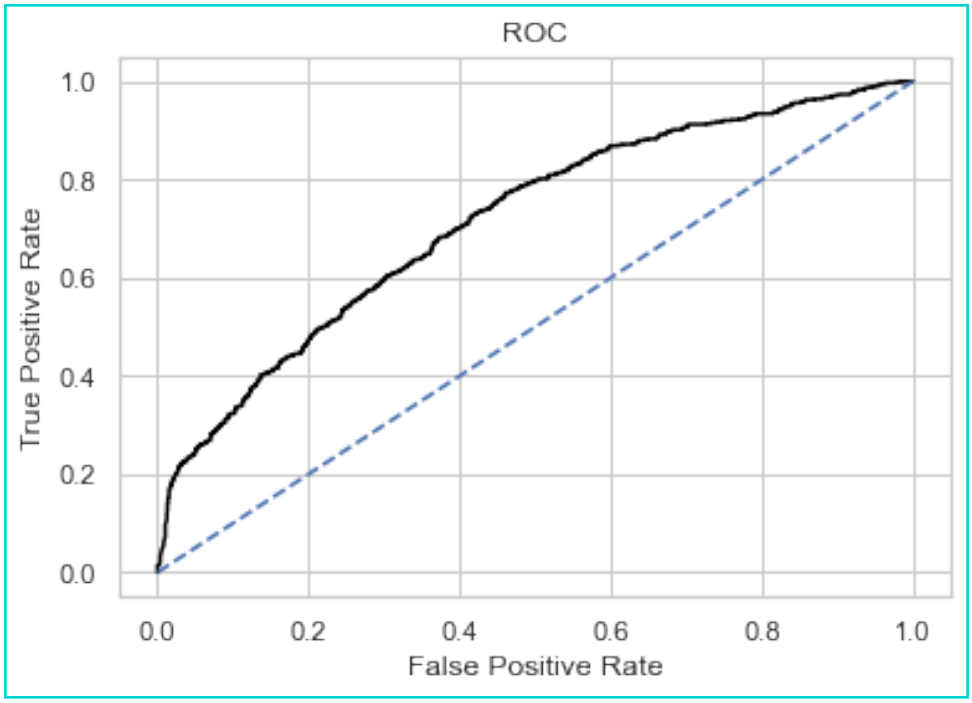



Figure No. 12.5

### Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 1215 | 145 |
| | Yes | 321 | 319 |

Table No. 12. 11

## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.89 | 0.84 | 1360 |
| 1 | 0.69 | 0.50 | 0.58 | 640 |
| accuracy | | | 0.77 | 2000 |
| macro avg | 0.74 | 0.70 | 0.71 | 2000 |
| weighted avg | 0.76 | 0.77 | 0.76 | 2000 |

Table No. 12.12

## Testing data

### Confusion Matrix

| Predicted Class | | True Class | |
|---|---|---|---|
| | | No | Yes |
| | No | 520 | 64 |
| | Yes | 135 | 139 |

Table No. 12. 13

## ROC Curve

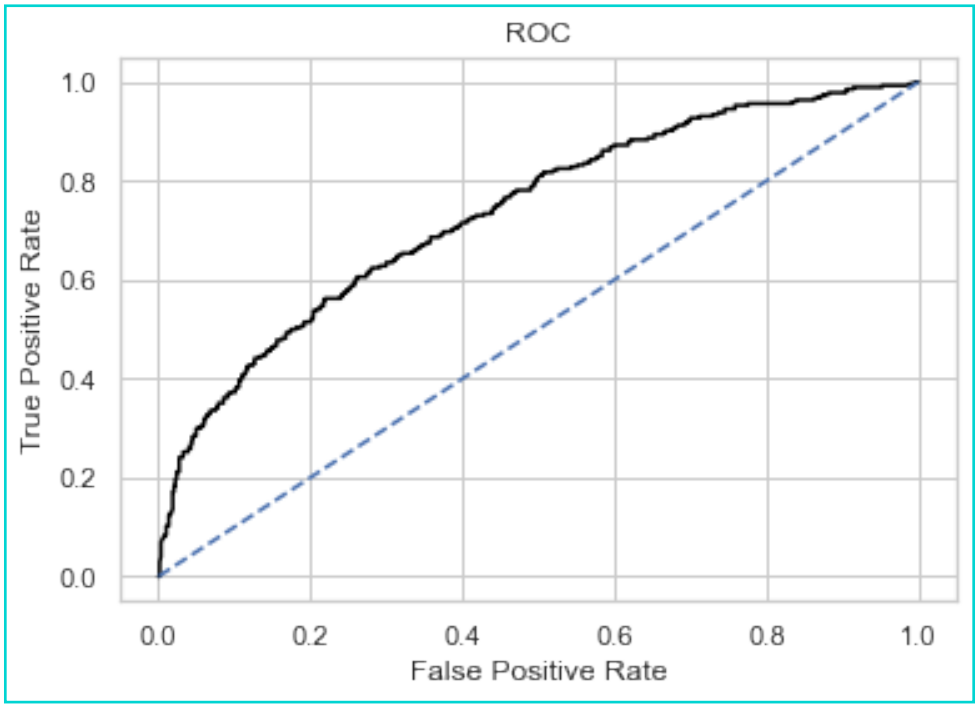



Figure No. 12.6

## Classification report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.79 | 0.89 | 0.84 | 584 |
| 1 | 0.68 | 0.51 | 0.58 | 274 |
| accuracy | | | 0.77 | 858 |
| macro avg | 0.74 | 0.70 | 0.71 | 858 |
| weighted avg | 0.76 | 0.77 | 0.76 | 858 |

Table No. 12.14

## ANN Conclusion

| | Train Data | Test Data |
|---|---|---|
| AUC | 62% | 62.9% |
| Accuracy | 89.3% | 76.8% |
| Sensitivity | 89.4% | 89.04% |
| Precision | 69% | 68% |
| Recall | 50% | 51% |
| f1-Score | 58% | 58% |

Table No. 12.15

Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

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

The Final Conclusion is as follows:

### 1. ROC Curve for the 3 models on the Training data

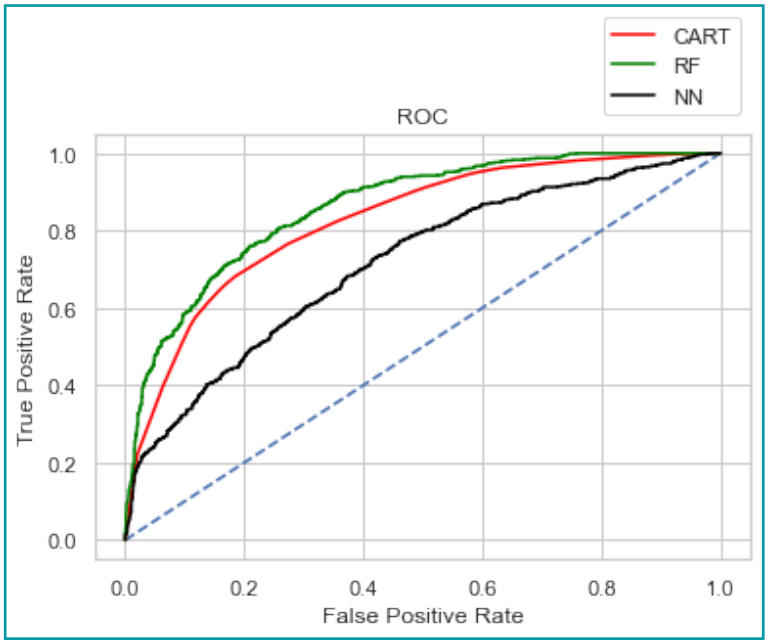
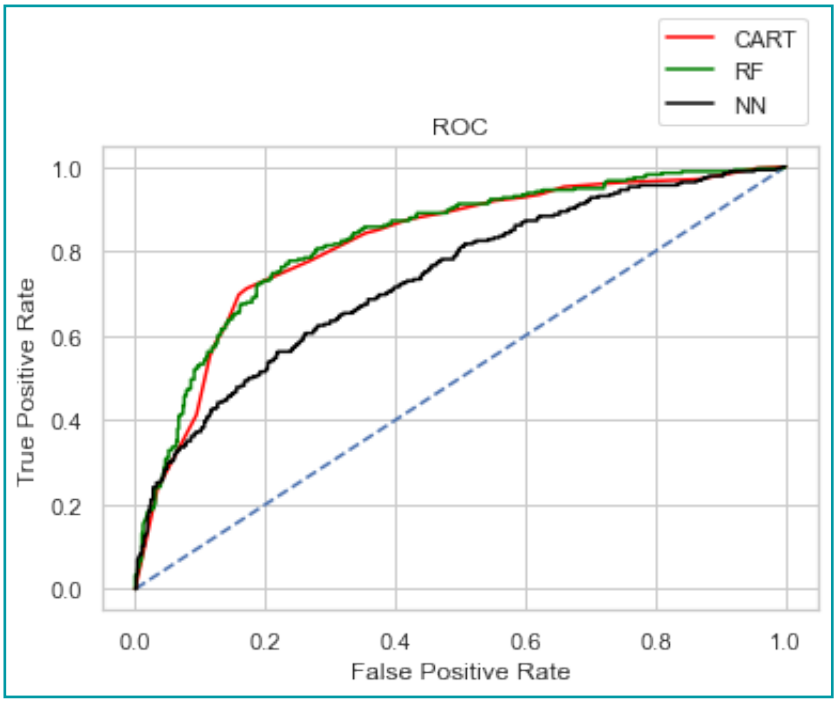



Figure No. 13

### 2. ROC Curve for the 3 models on the Test data



3. **Comparison of the performance metrics from the 3 models :**

| | CART Train | CART Test | Random Forest Train | Random Forest Test | Neural Network Train | Neural Network Test |
|---|---|---|---|---|---|---|
| **Accuracy** | 0.78 | 0.78 | 0.80 | 0.79 | 0.77 | 0.77 |
| **AUC** | 0.83 | 0.82 | 0.74 | 0.74 | 0.62 | 0.63 |
| **Recall** | 0.57 | 0.59 | 0.59 | 0.62 | 0.50 | 0.51 |
| **Precision** | 0.70 | 0.69 | 0.73 | 0.68 | 0.69 | 0.68 |
| **F1 Score** | 0.63 | 0.64 | 0.65 | 0.65 | 0.58 | 0.58 |

Table No. 13.1

According to above comparison data, following are results :

1. ROC Curve is better for RF and CART. Whereas ROC curve for NN is not so good.
2. Accuracy is approximately similar for all the three models.
3. Since, in our scenario, if the management is formed about less number of claims and more claims come, then false-negatives is a problem. Hence, **Recall** is a better measure than precision.
4. So, on the basis of this conclusion, Recall is slightly better in RF and CART.
5. Slight overfitting for RF model.

Hence, I will be selecting Random forest here.

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

Recommendations are as follows:

1. With the above analysis, we can find the claims expected from a customer or not and mitigate risk by counting the amount approximately , that is to be given to the customers.
2. Getting more customers, who has lesser chances of claiming.
3. Modifying plans to be offered to the customers, who has higher chances of claiming to reduce maximum amount going out from the Insurance company.