# DATA MINING

## PROJECT


### Abstract

Data mining is a field of intersection of computer science and statistics used to discover patterns in the information bank. The main aim of the data mining process is to extract the useful information from the dossier of data and mould it into an understandable structure for future use.



Sudheendra K

PGPDSBA.O.APR22.C

# Project Data Mining

## *Table of Contents*

# Project Data Mining

# Project Data Mining

## Problem 1: Clustering

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

*Data Dictionary 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).*

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

# Project Data Mining

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

TABLE 1. 1: DATA INFORMATION

- *As we can see the above table there are 210 rows with 7 columns where all the attributes are of float datatype and at first glance the data looks neat and clean.*
- *The shape of the data is (210,7).*
- *There are no anomalies or any kind of bad data present the dataframe.*
- *There are no Null Values present in the dataset.*

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

TABLE 1. 2*: HEAD OF THE DATAFRAME*

- *These are the top 5 rows of the data,*
- *with double digit values in Spending and advance_payments*
- *single digit values in current_balance, credit_limit, min_payment_amt and max_spent_in_single_shopping, and point values in probability_of_full_payment.*

### *1.1.1 Univariate Data Analysis*

*The analysis of univariate data is thus the simplest form of analysis since the information deals with only one quantity that changes. It does not deal with causes or relationships and the main purpose of the analysis is to describe the data and find patterns that exist within it.*

*Spending Attribute*

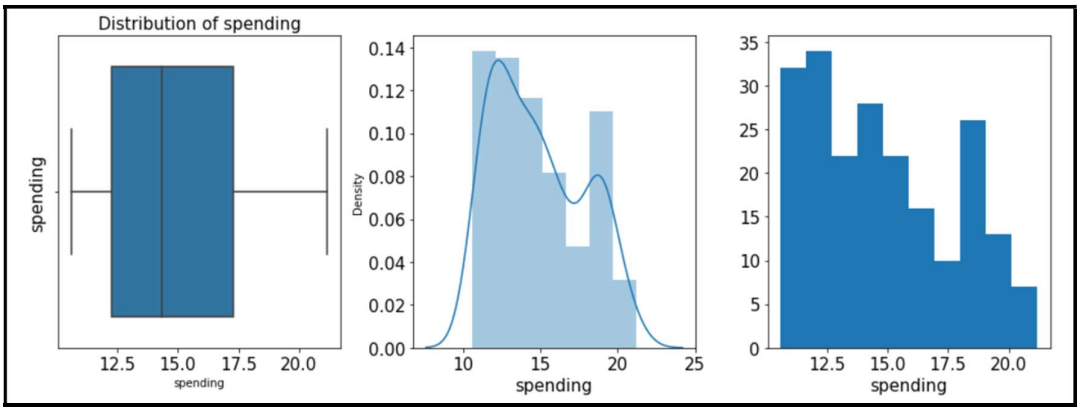


FIGURE 1. 1*: SPENDING ATTRIBUTE*

- *The 1st Quartile (Q1) for spending variable is: 12.27.*
- *The 3rd Quartile (Q3) for spending variable is: 17.305.*
- *Interquartile range (IQR) of spending is 5.035.*
- *Lower outliers in spending: 4.717499999999999.*
- *Upper outliers in spending: 24.8575.*
- *No outliers in the variable spending.*
- *This variable is right tailed skewed about 0.3998.*

*Advance Payments*

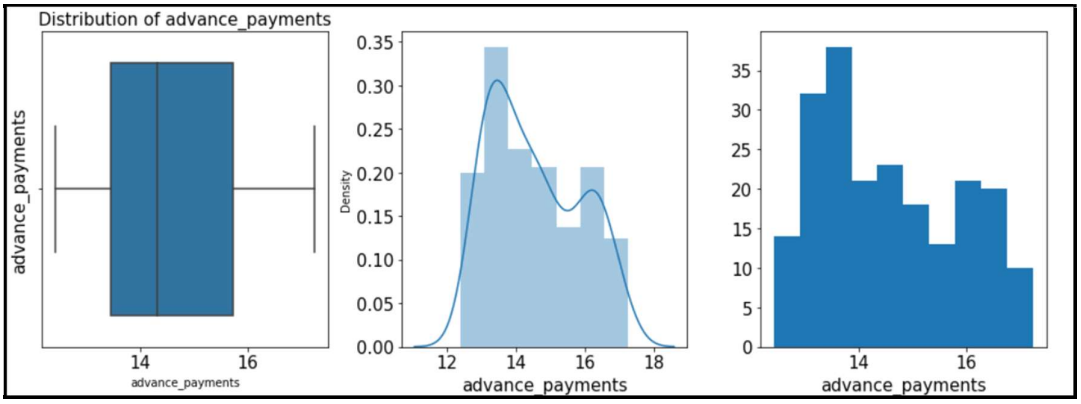


FIGURE 1. 2: *ADVANCE PAYMENTS ATTRIBUTE*

- *The 1st Quartile (Q1) for advance payment variable is: 13.45.*
- *The 3rd Quartile (Q3) for advance payment variable is: 15.715.*
- *Interquartile range (IQR) of advance payment is 2.265.*
- *Lower outliers in advance_payments: 10.052499999999998.*
- *Upper outliers in advance_payments: 19.1125.*
- *There are no outliers present in the variable advance payments.*
- *This variable is right tailed skewed about 0.3865.*

*Probability Of Full Payments*

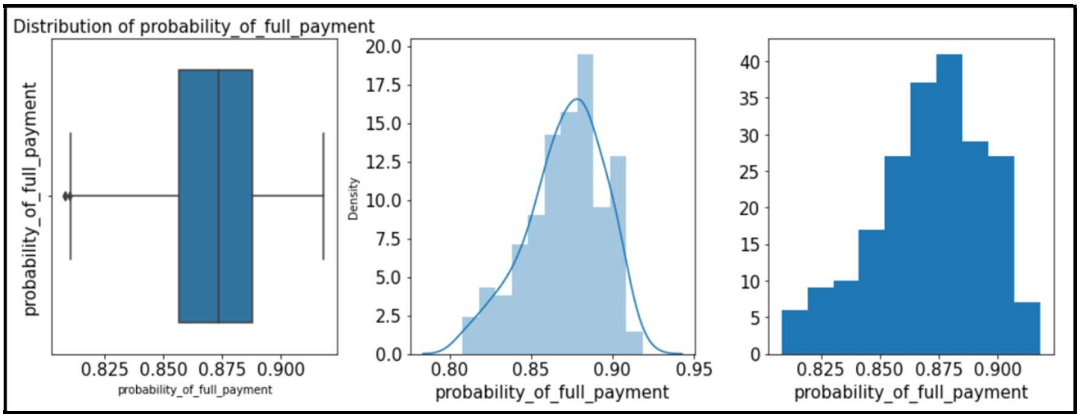


FIGURE 1. 3: *PROBABILITY OF FULL PAYMENTS - ATTRIBUTE*

- *The 1st Quartile (Q1) for Probability of full payments variable is: 0.8569.*
- *The 3rd Quartile (Q3) for Probability of full payments variable is: 0.8877.*
- *Interquartile range (IQR) of Probability of full payments is 0.0308.*

- *Lower outliers in Probability of full payments: 0.8105.*
- *Upper outliers in Probability of full payments: 0.9340.*
- *There are outliers present in the variable Probability of full payments in lower bound.*
- *If there are outliers present, they should be treated as clustering results are affected by the presence of outliers.*
- *This is the only variable which is left tailed skewed about -0.5379.*

*Current Balance*

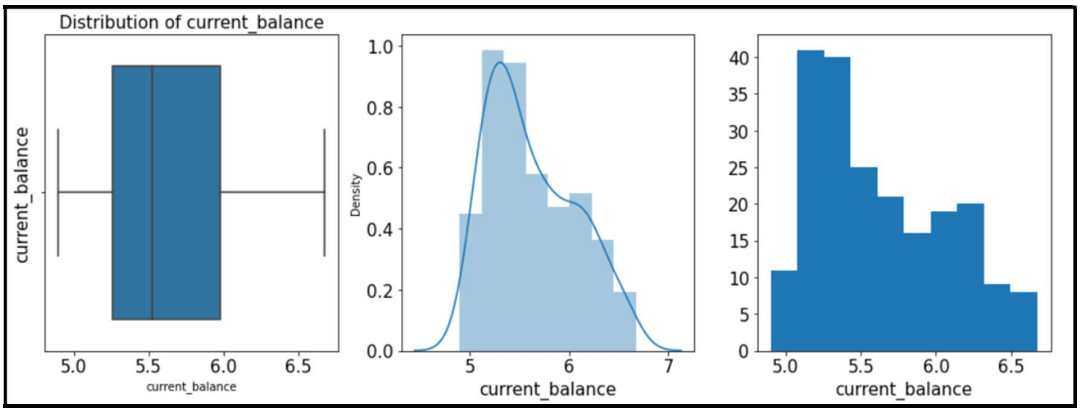


FIGURE 1. 4*: CURRENT BALANCE ATTRIBUTE*

- *The 1st Quartile (Q1) for Current Balance variable is: 5.2622.*
- *The 3rd Quartile (Q3) for Current Balance variable is: 5.9797.*
- *Interquartile range (IQR) of Current Balance is 0.7175.*
- *Lower outliers in Current Balance: 4.186.*
- *Upper outliers in Current Balance: 7.056.*
- *There are no outliers present in the variable Current Balance.*
- *This variable is right tailed skewed about 0.5254.*

*Credit Limit*

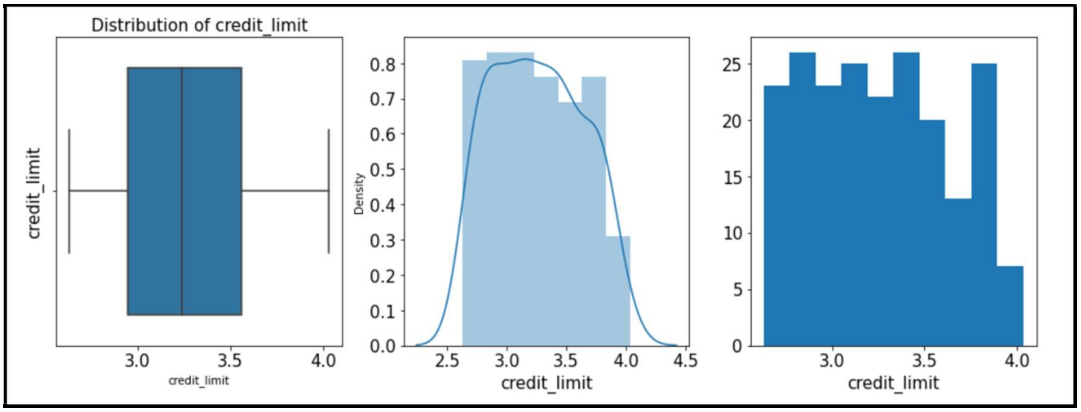


FIGURE 1. 5*: CREDIT LIMIT ATTRIBUTE*

- *The 1st Quartile (Q1) for Credit Limit variable is: 2.94.*
- *The 3rd Quartile (Q3) for Credit Limit variable is: 3.56.*
- *Interquartile range (IQR) of Credit Limit is 0.6177.*
- *Lower outliers in Credit Limit: 2.944.*
- *Upper outliers in Credit Limit: 3.5617.*
- *There are no outliers present in the variable Credit Limit.*
- *This variable is right tailed skewed about 0.1343.*

*Min Payment Amount*

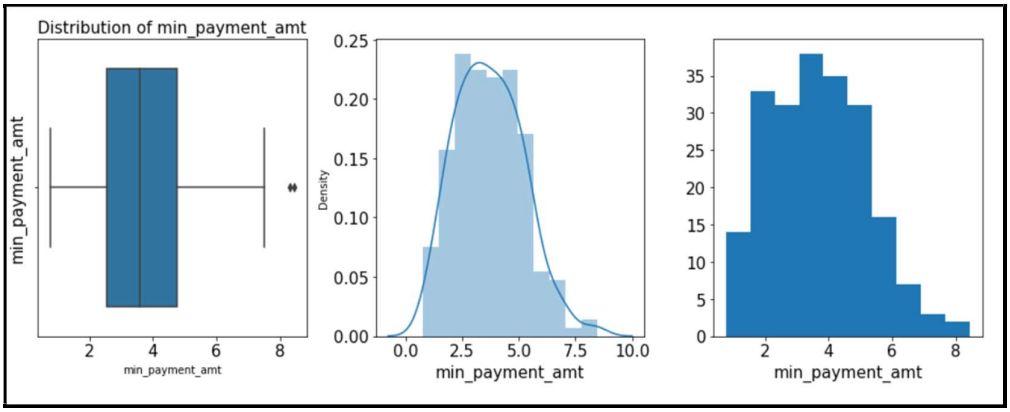


FIGURE 1. 6*: MIN PAYMENT AMOUNT ATTRIBUTE*

- *The 1st Quartile (Q1) for Min Payment Amount variable is: 2.5615.*
- *The 3rd Quartile (Q3) for Min Payment Amount variable is: 4.7687.*
- *Interquartile range (IQR) of Min Payment Amount is 2.2072.*
- *Lower outliers in Min Payment Amount: -0.7493.*

- *Upper outliers in Min Payment Amount:  8.07962.*
- *There are outliers present in the variable Min Payment Amount in upper bound.*
- *If there are outliers present, they should be treated as clustering results are affected by the presence of outliers.*
- *This variable is right tailed skewed about 0.4016.*

*Max Spent In Single Shopping*

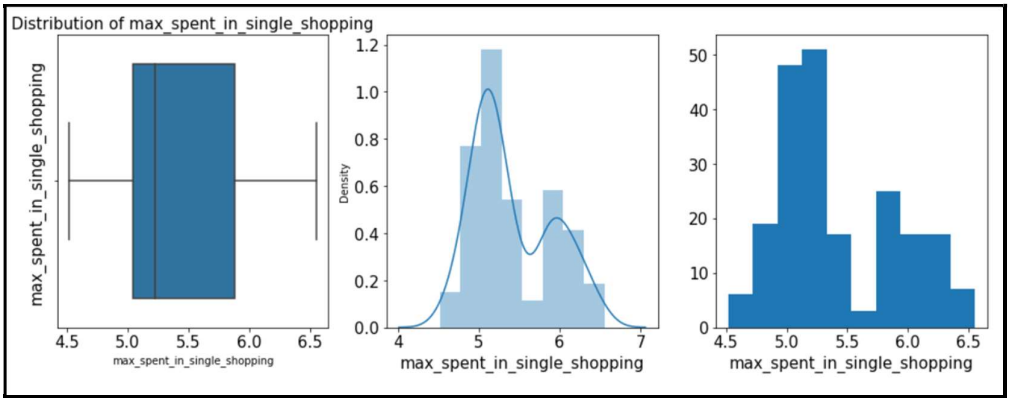


FIGURE 1. 7*: MAX SPENT IN SINGLE SHOPPING*

- *The 1st Quartile (Q1) for Max Spent in Single Shopping is:  5.045.*
- *The 3rd Quartile (Q3) for Max Spent in Single Shopping is:  5.877.*
- *Interquartile range (IQR) of Max Spent in Single Shopping is 0.8319.*
- *Lower outliers in Max Spent in Single Shopping:  3.797.*
- *Upper outliers in Max Spent in Single Shopping:  7.125.*
- *There are no outliers present in the variable Max Spent in Single Shopping.*
- *This variable is right tailed skewed about 0.5618.*

***Kindly Note: Though I did treat the outlier, there can be still seen one as per the boxplot, I consider this is okay, as no extremes outliers found in both the above said variables.***

### *1.1.2 Multivariate Analysis*

- *Checking for multicollinearity.*

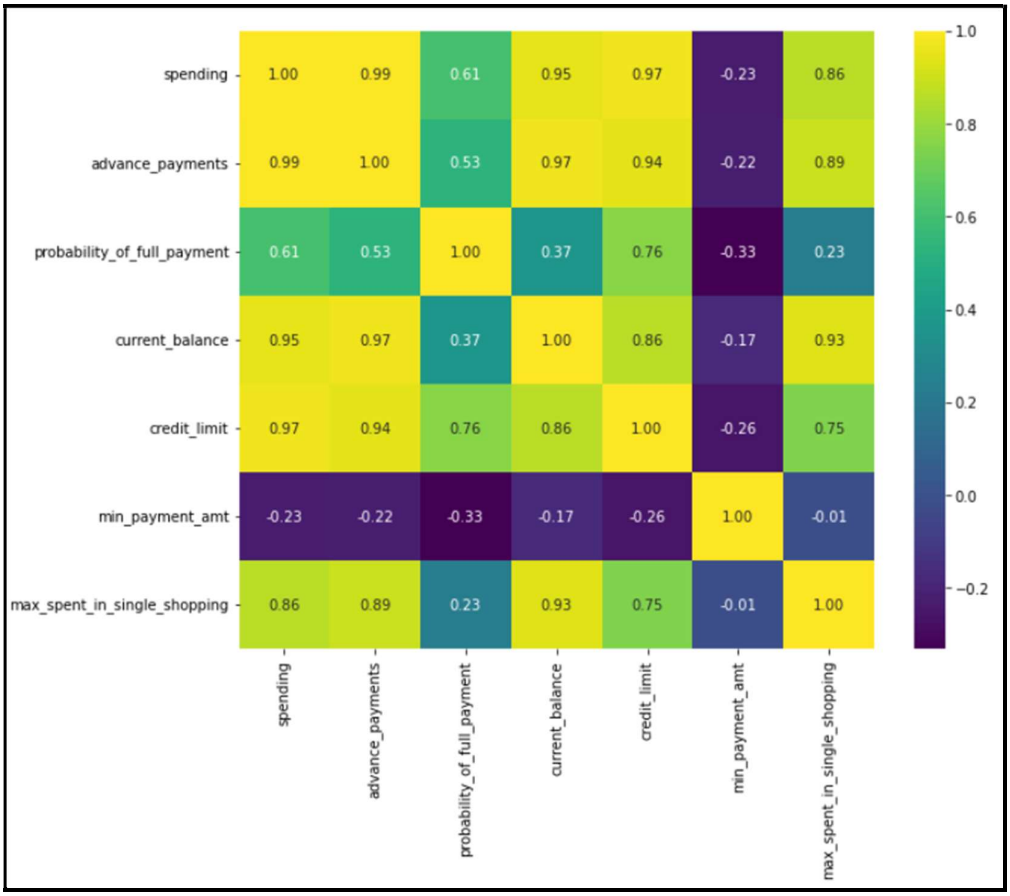


FIGURE 1. 8*: HEAT MAP*

| | spending | advance_payments | probability_of_full_payment | current_balance | credit_limit | min_payment_amt | max_spent_in_single_shopping |
|---|---|---|---|---|---|---|---|
| spending | 1.000000 | 0.994341 | 0.608288 | 0.949985 | 0.970771 | -0.229572 | 0.863693 |
| advance_payments | 0.994341 | 1.000000 | 0.529244 | 0.972422 | 0.944829 | -0.217340 | 0.890784 |
| probability_of_full_payment | 0.608288 | 0.529244 | 1.000000 | 0.367915 | 0.761635 | -0.331471 | 0.226825 |
| current_balance | 0.949985 | 0.972422 | 0.367915 | 1.000000 | 0.860415 | -0.171562 | 0.932806 |
| credit_limit | 0.970771 | 0.944829 | 0.761635 | 0.860415 | 1.000000 | -0.258037 | 0.749131 |
| min_payment_amt | -0.229572 | -0.217340 | -0.331471 | -0.171562 | -0.258037 | 1.000000 | -0.011079 |
| max_spent_in_single_shopping | 0.863693 | 0.890784 | 0.226825 | 0.932806 | 0.749131 | -0.011079 | 1.000000 |

TABLE 1. 3*: CORRELATION TABLE*

*Strong positive correlation between the below variables*

- *spending & advance_payments..*
- *advance_payments & current_balance.*
- *credit_limit & spending.*
- *spending & current_balance.*
- *credit_limit & advance_payments.*

- *max_spent_in_single_shopping & current_balance.*

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

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| spending | 210.0 | 14.847524 | 2.909699 | 10.5900 | 12.27000 | 14.35500 | 17.305000 | 21.1800 |
| advance_payments | 210.0 | 14.559286 | 1.305959 | 12.4100 | 13.45000 | 14.32000 | 15.715000 | 17.2500 |
| probability_of_full_payment | 210.0 | 0.870999 | 0.023629 | 0.8081 | 0.85690 | 0.87345 | 0.887775 | 0.9183 |
| current_balance | 210.0 | 5.628533 | 0.443063 | 4.8990 | 5.26225 | 5.52350 | 5.979750 | 6.6750 |
| credit_limit | 210.0 | 3.258605 | 0.377714 | 2.6300 | 2.94400 | 3.23700 | 3.561750 | 4.0330 |
| min_payment_amt | 210.0 | 3.700201 | 1.503557 | 0.7651 | 2.56150 | 3.59900 | 4.768750 | 8.4560 |
| max_spent_in_single_shopping | 210.0 | 5.408071 | 0.491480 | 4.5190 | 5.04500 | 5.22300 | 5.877000 | 6.5500 |

TABLE 1. 4*: DESCRIPTION OF THE ATTRIBUTES*

### *1.2.1 Scaling*

- *Well Yes, Scaling is necessary for clustering in this case as standardising the data prevents variables with larger scales from dominating the clustering process as the values of the variable's weightage may hugely vary, spending, advance payments are in different values and this may get more weightage.*

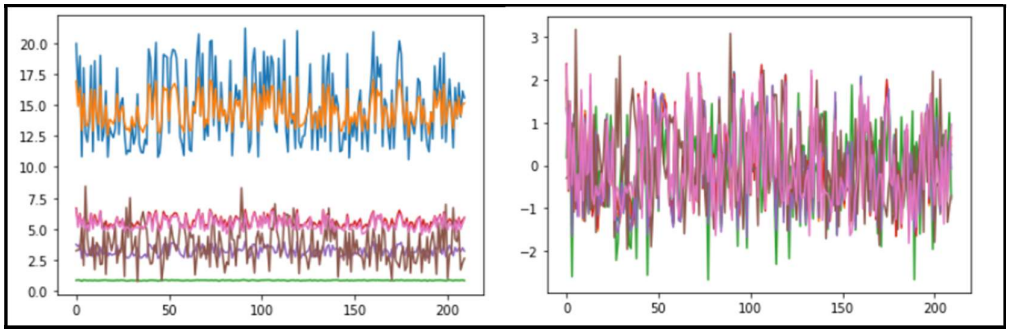


FIGURE 1. 9*: BEFORE SCALING & AFTER SCALING*

- *From the above plots, we can clearly notice that applying Z-score in our dataset has generated major standard deviations. It implies the data is more concentrated around the mean by doing scaling.*

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

TABLE 1. 5*: HEAD OF THE DATA FRAME AFTER SCALING*

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

### *1.3.1 Hierarchical Clustering*

- *We construct a Dendrogram for the scaled data and we obtain different cluster patterns using different linkages and distance criterions. Ward method was chosen after analysing all the dendrograms:*

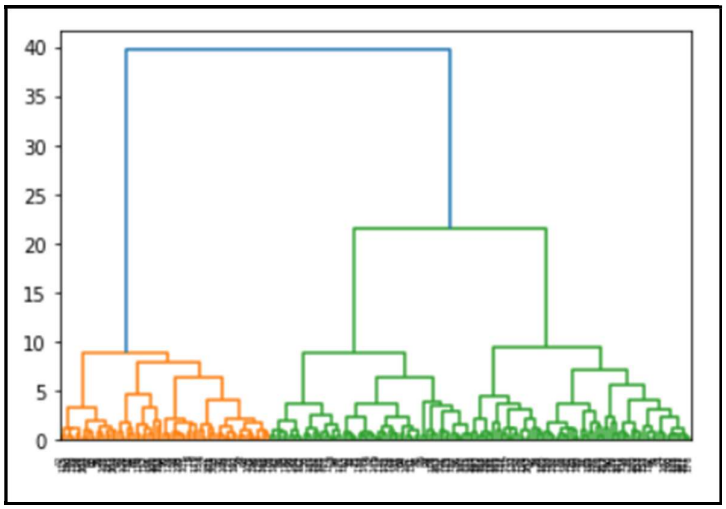


FIGURE 1. 10*: DENDOGRAM (WARD LINKAGE METHOD)*

- *Let's have a closer visualization,.*

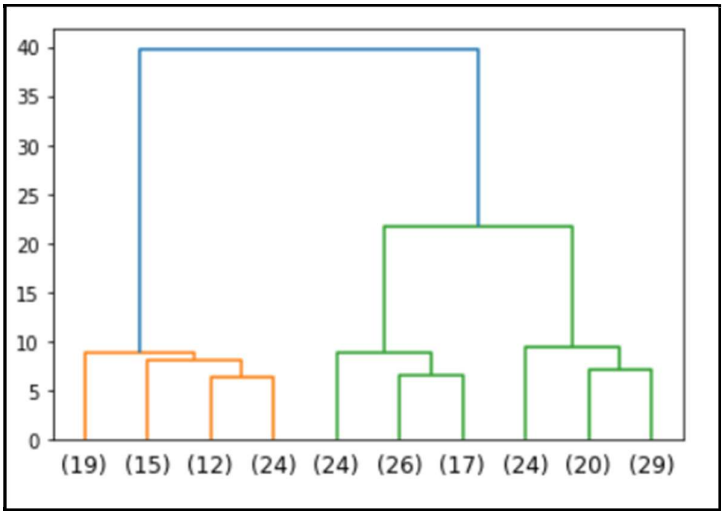


FIGURE 1. 11*: LAST 10- DENDOGRAM*

- *Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other.*

- *Upon performing hierarchical clustering on scaled dataset, we obtain mean values within 3 cluster formations as follows:*

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

TABLE 1. 6*: 3 CLUSTERS FORMATIONS*

- *Both methods are almost having similar means, minor variation, which we know it occurs, cluster grouping based on the dendrogram, 3 or max 4 looks good. Did the further analysis, and based on the dataset had gone for 3 group cluster solution with based on using hierarchical clustering.*

- *Also in real time, there could have been more variables value captured like tenure, balance, purchase, instalment of purchase, others.*

- *And three group cluster solution gives a pattern based on high/medium/low spending with max spent in single shopping (high value item) and probability of full payment (payment made).*

## 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.

### *1.4.1 K – Means Clustering*

- *K-Means is a non-hierarchical approach to forming good clusters is to prespecify a desired number of clusters, 'K'.*

- *The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.*

- *In this method the partitions are made such that non-overlapping groups having no hierarchical relationships between themselves.*

- *K-means clustering is widely used in large dataset applications.*

- *Using SKlearn's K-Means package, we fit the scaled data, calculating the inertia and then total within-cluster sum of squares (WSS).*

- *That value of 'K' is chosen to be optimum, where addition of one more cluster does not lower the value of total sum of squares appreciably.*

*1.4.2 Elbow Curve*

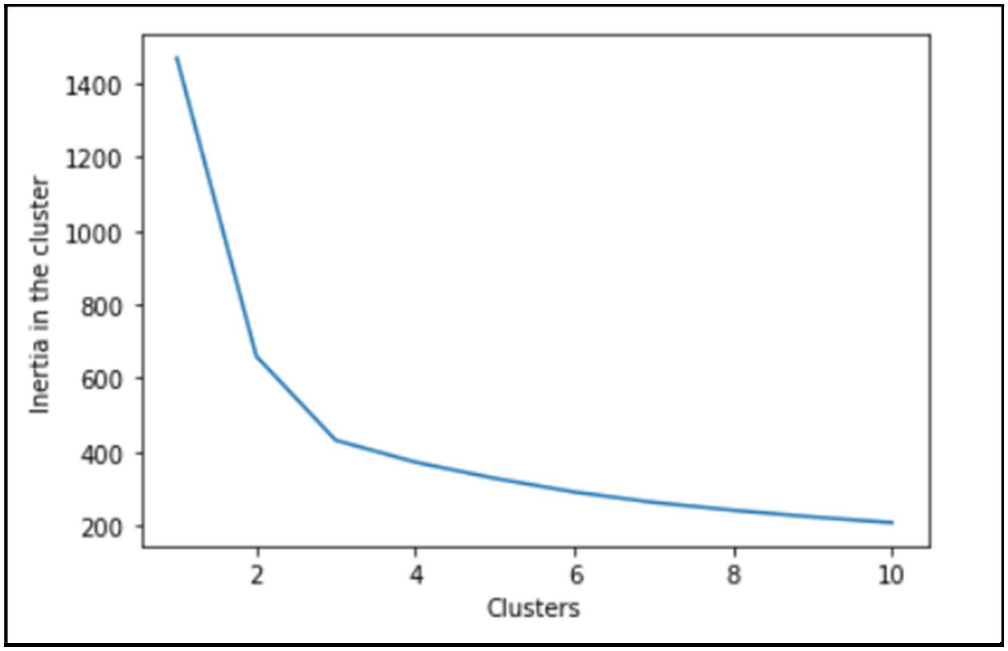


FIGURE 1. 12*: ELBOW CURVE*

- *Now with the help of Elbow Curve, for a given number of clusters, the total within-cluster sum of squares is computed.*

### *1.4.3 Silhouette Score*

- *This method measures how tightly the observations are clustered and also the average distance between clusters.*

- *For each observation a silhouette score is constructed which is the function of average distance between the point and all other points in the cluster to which it belongs, and the distance between the point and all other points in all other clusters, that it does not belong to. The maximum value of the statistic indicates the optimum value of 'K'.*

- *Upon performing Non-Hierarchical clustering on scaled dataset, we obtain mean values within 3 cluster formations as follows:*

| cluster | 1 | 2 | 3 |
|---|---|---|---|
| spending | 14.4 | 11.9 | 18.5 |
| advance_payments | 14.3 | 13.2 | 16.2 |
| probability_of_full_payment | 0.9 | 0.8 | 0.9 |
| current_balance | 5.5 | 5.2 | 6.2 |
| credit_limit | 3.3 | 2.8 | 3.7 |
| min_payment_amt | 2.7 | 4.7 | 3.6 |
| max_spent_in_single_shopping | 5.1 | 5.1 | 6.0 |

TABLE 1. 7*: SILHOUETTE SCORE*

- *However, the Silhouette Score of 2 clusters was more appropriate, but of this clustering effort is to devise a suitable recommendation system.*

- *It may not be practical to manage a very low number of tailor-made recommendations. Therefore, Cluster number = 3 serves the purpose of our requirement to produce valuable insights.*

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

*Cluster Group Profiles*

- *Group 1: High Spending.*
- *Group 3: Medium Spending.*
- *Group 2: Low Spending.*

*Recommendations:*

### *1.5.1 Group 1: High Spending Group*

- *Offerings higher reward points on a higher purchase could increase their spending capacity.*
- *Adding an option of no cost EMI as a promotional scheme with bank's tied up brands, can be a great motivator for this group.*

- *The segmentation of maximum max spent in single shopping is the highest of this group, hence, the discounts offered or attractive offers on the next transactions with full payments upfront.*

- *Periodic assessment and increase of credit limits.*

- *The preferential customer treatment which might lead to higher spending habits.*

- *Since there is a clear indication that the customers of this category are financially stable, interesting loan schemes exclusively for them could be planned.*

- *Collaborations with high end luxury brands and accessories would lead to higher one-time maximum spending.*

### *1.5.2 Group 3: Medium Spending Group*

- *The customers of this segmentation cluster are suggested to be the target customers with highest potential as there is consistent maintenance of a higher credit score which results in timely payments of their bills.*

- *The customers of this category can have an increased credit limit raised and monitored periodically and have significantly marginalised interest rates keeping RBI guidelines in mind.*

- *The advertisement and promotion of premium cards or loyalty cards of specific brand collaborated partnerships would lead to increase in the transactional values over an extended period.*

- *Once the above-mentioned credit limits are enhanced, the result would be an automatic increase in spending habits across the premium partners in e-commerce, travel portals, airlines & hotels.*

### *1.5.3 Group 2: Low Spending Group*

- *We can spend some time analysing the brands and utilities this segment spends its most amount on and provide discounts and offers on the credit card usage accordingly.*

- *Customers of this segment will have to be given timely reminders on payments so that the due dates of the billing cycles are not missed.*

- *Small-scale campaigns could be run providing the customers of this segment attractive offers for early payments which would improve the rate of payment received and result in lesser default rates.*

**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.**

*Attribute Information:*

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

# Project Data Mining

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

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

TABLE 2. 1*: HEAD OF THE DATAFRAME*

- *After importing the necessary libraries and data in the python notebook, above is the top 5 rows of the data.*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   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   Commision     3000 non-null   float64
 5   Channel       3000 non-null   object 
 6   Duration      3000 non-null   int64  
 7   Sales         3000 non-null   float64
 8   Product Name  3000 non-null   object 
 9   Destination   3000 non-null   object 
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

TABLE 2. 2*: DATA INFORMATION*

- *In the above image, we can see that there are no null values in the data.*

- *2 of the 10 variables are of Data type are Float, 2 variables are of Data type are Integer and the remaining 6 are of Object Data type.*

- *The shape of the data is (3000, 10).*

| | Age | Agency_Code | Type | Claimed | Commision | 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 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **2940** | 36 | EPX | Travel Agency | No | 0.0 | Online | 8 | 10.0 | Cancellation Plan | ASIA |
| **2947** | 36 | EPX | Travel Agency | No | 0.0 | Online | 10 | 28.0 | Customised Plan | ASIA |
| **2952** | 36 | EPX | Travel Agency | No | 0.0 | Online | 2 | 10.0 | Cancellation Plan | ASIA |
| **2962** | 36 | EPX | Travel Agency | No | 0.0 | Online | 4 | 20.0 | Customised Plan | ASIA |
| **2984** | 36 | EPX | Travel Agency | No | 0.0 | Online | 1 | 20.0 | Customised Plan | ASIA |

139 rows × 10 columns

TABLE 2. 3: DUPLICATES

- *Though it shows there are 139 records, but it can be of different customers, there is no customer ID or any unique identifier, so I am not dropping them off.*

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

TABLE 2. 4*: DESCRIPTION OF THE DATA*

- *For Object data type variables like, Agency code, Type, Claimed, Channel, Product Name, and Destination, there are very less 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 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.*

- *There is a negative value present in the duration variable seems like a wrongly entered data.*

### *2.1.1 Univariate Analysis*

*Age Variable*

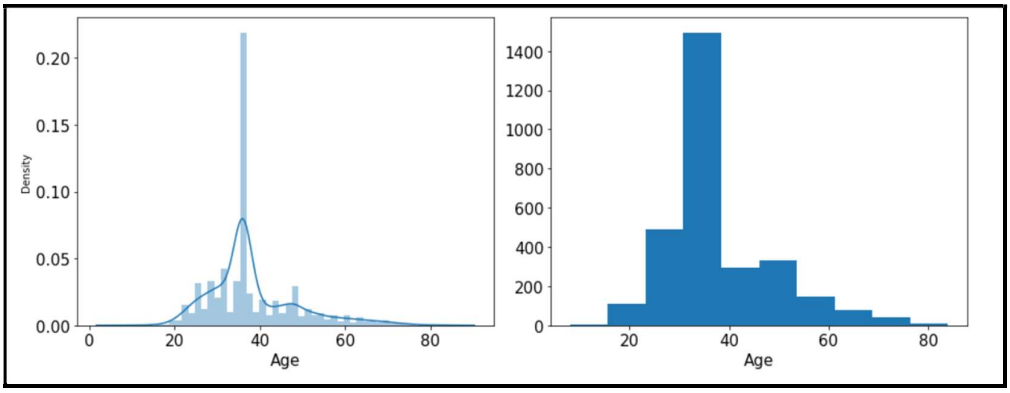


FIGURE 2. 1*: DIST PLOT & HISTOGRAM OF AGE*

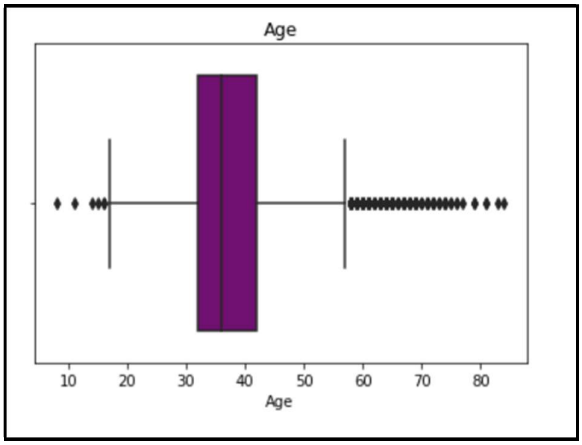


FIGURE 2. 2*: BOX PLOT FOR AGE VARIABLE*

- *The 1st Quartile (Q1) for Age is: 32.*
- *The 3rd Quartile (Q3) for Age is: 43.*
- *Interquartile range (IQR) of Age is 10.*
- *Lower outliers in Age: 17.*
- *Upper outliers in Age: 24.8575.*
- *There are outliers present in the variable Age.*
- *The range of the variable Age is 76.*

*Commission Variable*

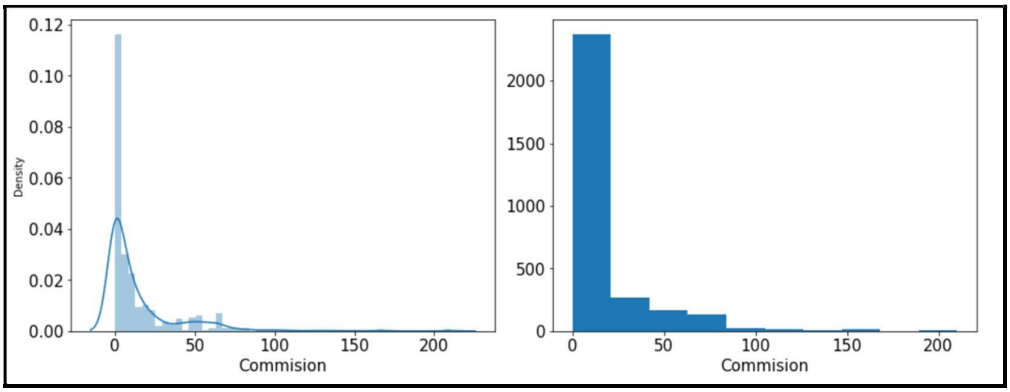


FIGURE 2. 3*: DISTPLOT & HISTOGRAM OF COMMISSION VARIABLE*

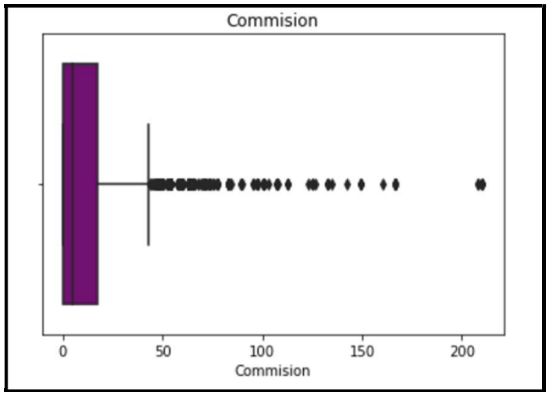


FIGURE 2. 4*: BOX PLOT FOR COMMISSION VARIABLE*

- *The 1st Quartile (Q1) for Commission is: 0.*
- *The 3rd Quartile (Q3) for Commission is: 17.235.*
- *Interquartile range (IQR) of Commission is 17.235.*
- *Lower outliers in Commission: -25.85.*
- *Upper outliers in Commission: 43.08.*
- *There are outliers present in the variable Commission.*
- *The range of the variable Age is 210.*

*Duration Variable*

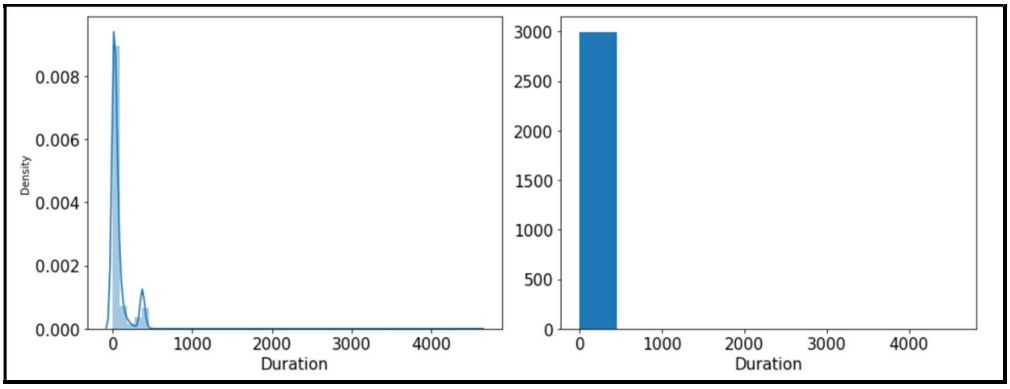


FIGURE 2. 5*: DISTPLOT & HISTOGRAM OF DURATION VARIABLE*

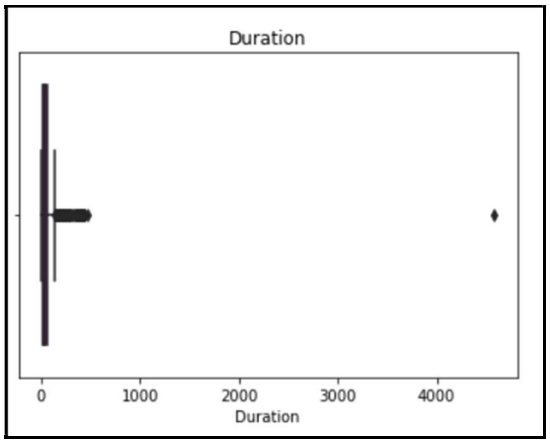


FIGURE 2. 6*: BOX PLOT FOR DURATION*

- *The 1st Quartile (Q1) for Commission is: 11.*
- *The 3rd Quartile (Q3) for Commission is: 63.*
- *Interquartile range (IQR) of Commission is 52.*
- *Lower outliers in Commission: -67.*
- *Upper outliers in Commission: 141.*
- *There are outliers present in the variable Commission.*
- *The range of the variable Commission is 4581.*

*Sales Variable*

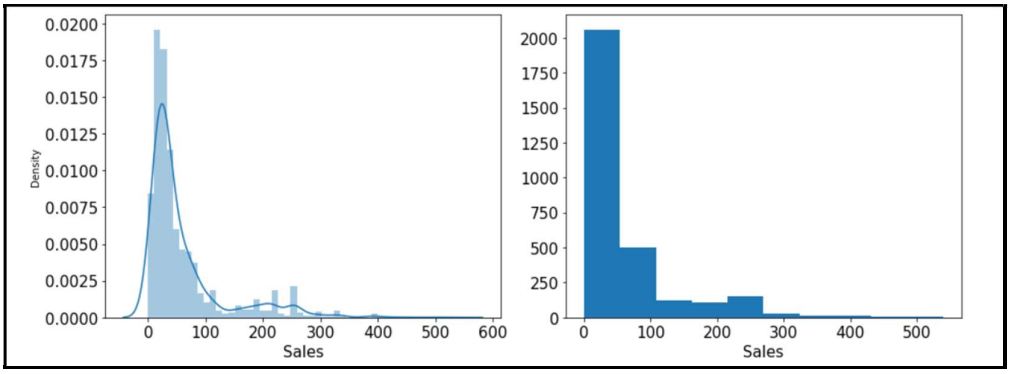


FIGURE 2. 7*: DISTPLOT & HISTOGRAM OF SALES VARIABLE*

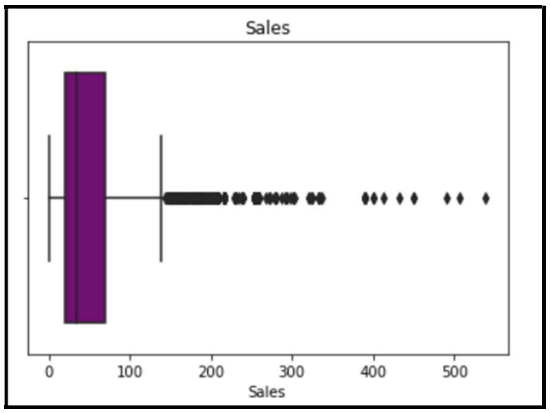


FIGURE 2. 8*: BOX PLOT FOR SALES VARIABLE*

- *The 1st Quartile (Q1) for Sales is: 11.*
- *The 3rd Quartile (Q3) for Sales is: 63.*
- *Interquartile range (IQR) of Sales is 52.*
- *Lower outliers in Sales: -67.*
- *Upper outliers in Sales: 141.*
- *There are outliers present in the variable Sales.*
- *The range of the variable Sales is 4581.*

*Type Variable*

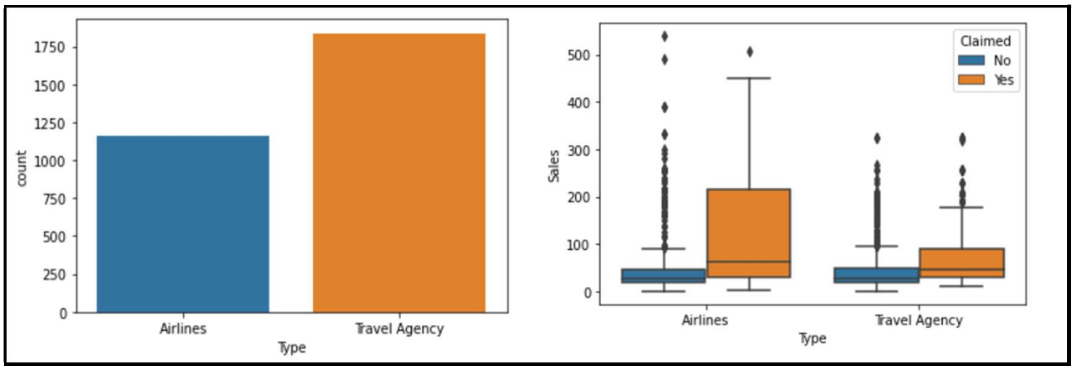


FIGURE 2. 9*: COUNTPLOT & BOX FOR TYPE VARIABLE*

- *This variable has two features namely Airlines and Travel Agency.*
- *We can observe that highest counts are from Travel agency and we could also notice through airlines seems to averagely done.*
- *There are outliers present in the variable Type.*

*Channel Variable*

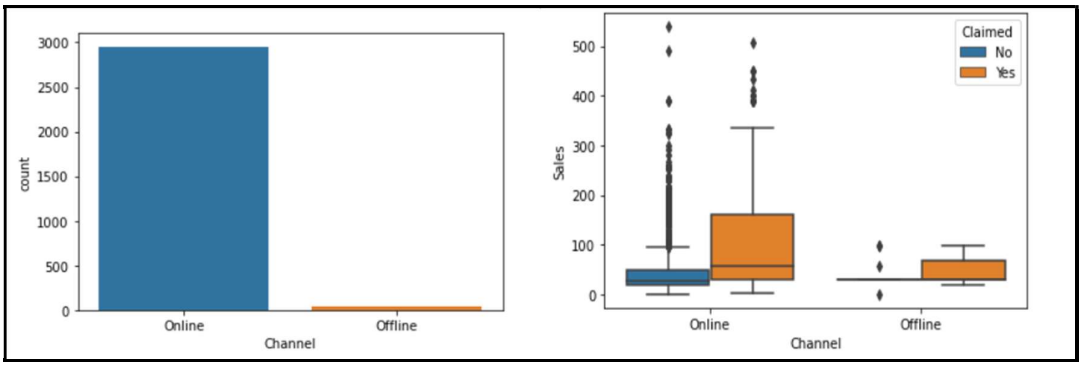


FIGURE 2. 10*: COUNTPLOT & BOX FOR CHANNEL VARIABLE*

- *This feature has two modes namely Online and offline.*
- *We can observe that highest frequency of sales is done through online channel.*
- *There are outliers present in the variable Channel.*

*Product Name*

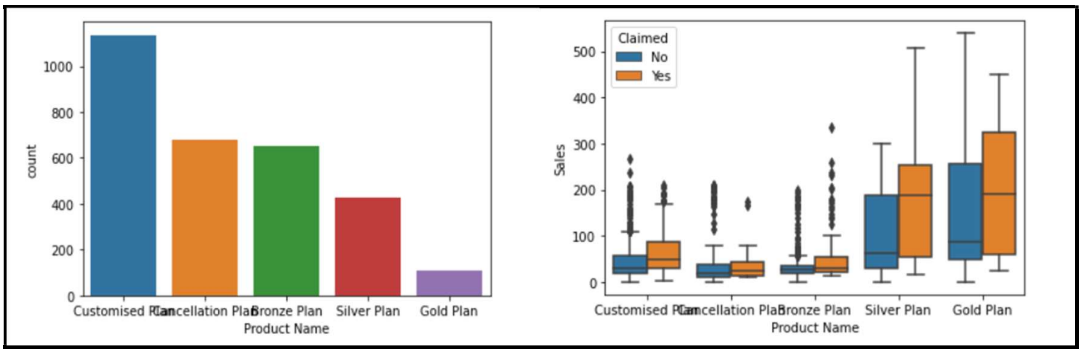


FIGURE 2. 11*: COUNTPLOT & BOX FOR PRODUCT NAME*

- *From the above figure we can notice that this feature has five objects Customised Plan, Cancellation Plan, Bronze Plan, Silver Plan and Gold plan.*
- *The Highest plan is for Customised insurance and the lowest is for Gold Loan can be seen in the variable Product Name.*
- *There are outliers present in the variable Product.*

*Destination Variable*

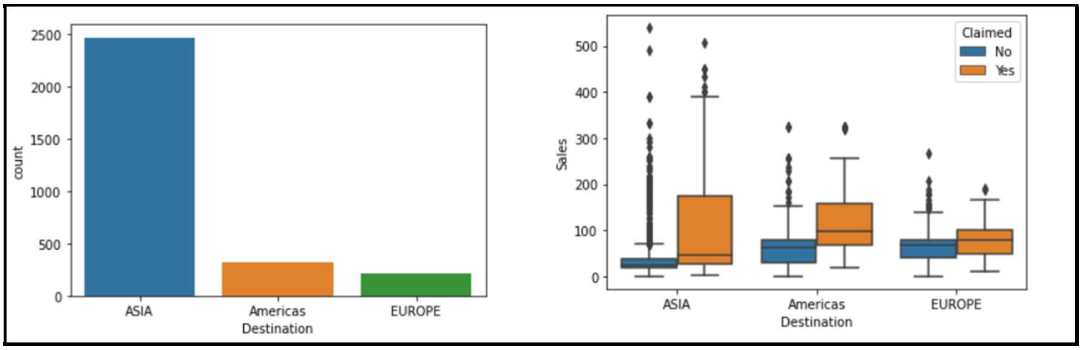


FIGURE 2. 12*: COUNTPLOT & BOX FOR DESTINATION VARIABLE*

- *This variable consists of three destinations they are, Asia, America and Europe.*
- *Asia stands in the first place and Europe stands at the last*
- *There are outliers in the variable Destination.*

### *2.1.2 Multivariate Analysis*

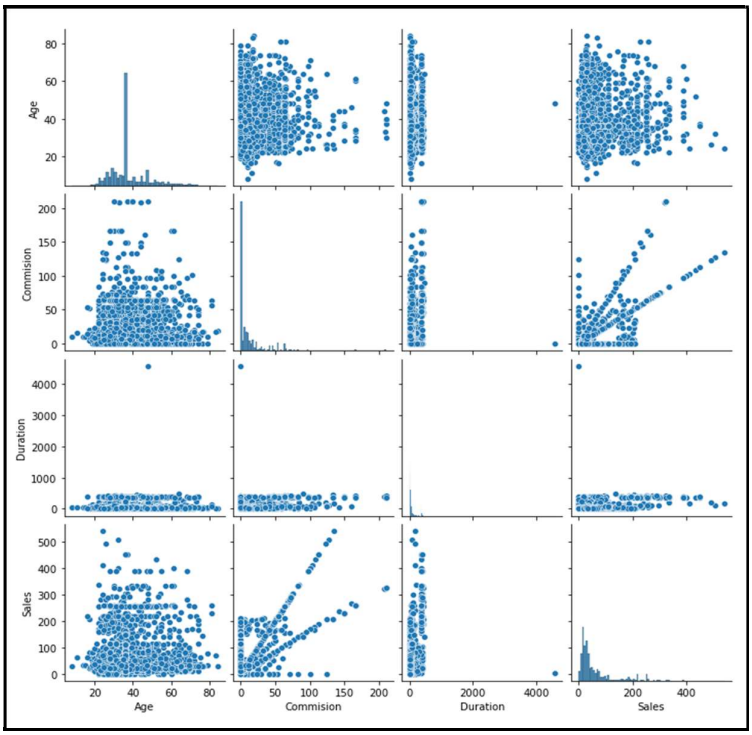


FIGURE 2. 13*: PAIR PLOT OF QUANTITATIVE ATTRIBUTES*

- *From the above figure pair plot and the below figure heat map, we can observe that there is no major correlation in any of the two variables In comparison, between Sales and Commission has a positive correlation of 0.76 which is high in comparison with other variables.*

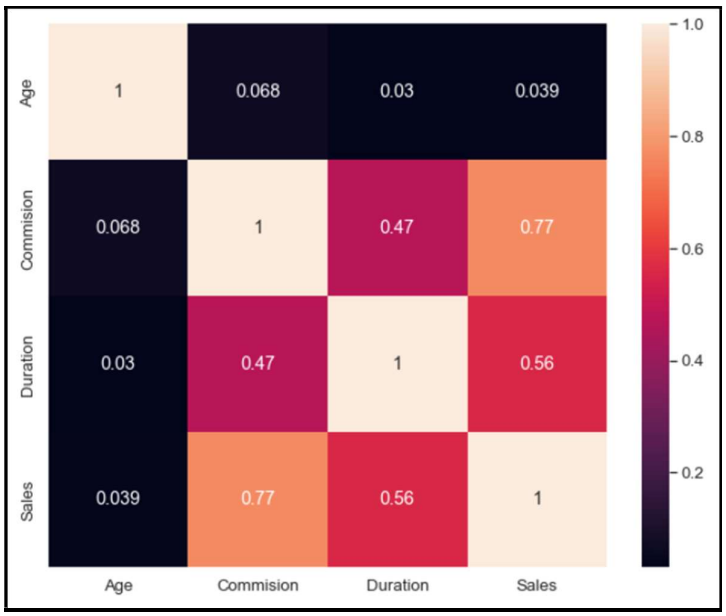


FIGURE 2. 14*: HEAT MAP*

### *2.1.3 Encoding*

- *On the next step, we are changing the data type of Object variables into Categorical data. After which, the all the data types of the data are either Integer or Float which can noticed in the below table:*

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   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   Commision      3000 non-null   float64
 5   Channel        3000 non-null   int8
 6   Duration       3000 non-null   int64
 7   Sales          3000 non-null   float64
 8   Product Name   3000 non-null   int8
 9   Destination    3000 non-null   int8
dtypes: float64(2), int64(2), int8(6)
memory usage: 111.5 KB
```

TABLE 2. 5*: DATA INFORMATION AFTER ENCODING*

| | Age | Agency_Code | Type | Claimed | Commision | Channel | Duration | Sales | Product Name | Destination |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 48 | 0 | 0 | 0 | 0.70 | 1 | 7 | 2.51 | 2 | 0 |
| **1** | 36 | 2 | 1 | 0 | 0.00 | 1 | 34 | 20.00 | 2 | 0 |
| **2** | 39 | 1 | 1 | 0 | 5.94 | 1 | 3 | 9.90 | 2 | 1 |
| **3** | 36 | 2 | 1 | 0 | 0.00 | 1 | 4 | 26.00 | 1 | 0 |
| **4** | 33 | 3 | 0 | 0 | 6.30 | 1 | 53 | 18.00 | 0 | 0 |

TABLE 2. 6*: HEAD OF THE DATA AFTER ENCODING*

- *After Encoding the datasets, the glance could be observed from the above table.*

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

*Initially, splitting the data into Train and Test data.*

```
X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
```

SNIPPET 2. 1*: SHAPE AFTER SPLIT*

- *From the above snippet we can see the shape of Train and Test data after splitting the dependent and independent variable.*

### *2.2.1 CART MODEL*

- *CART is a Binary Decision Tree model. I have used Gini Index as its Criteria. It is an attribute that Maximizes the reduction in impurity is chosen as the Splitting Attribute.*

- *Using the Decision Tree Classifier and the Grid search method with multiple combinations of iterations, below are the best grid results:*

    - *'criterion': 'gini',*
    - *'max_depth': 4,*
    - *'min_samples_leaf': 25,*
    - *'min_samples_split': 300*

- *After looking at the decision tree, you can see the extraction of the variable importance shown below:*

| | Imp |
|---|---|
| Agency_Code | 0.625147 |
| Sales | 0.259036 |
| Product Name | 0.079851 |
| Commision | 0.023525 |
| Duration | 0.012441 |
| Age | 0.000000 |
| Type | 0.000000 |
| Channel | 0.000000 |
| Destination | 0.000000 |

SNIPPET 2. 2*: IMPORTANT VARIABLE AS PER CART*

- *As per the above extract, Agency code is being the most important variable in the dataset, followed by Sales and Product Name.*

- *Commission has comparatively very lesser significance, however Age, Type, Channel, Duration and Destination have least importance in the model building.*

### *2.2.2 RANDOM FOREST*

- *Random Forest Consists of many individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction. Class with most votes becomes model's prediction.*

- *Using the random forest classifier and grid search function running multiple combinations of iterations, below are the best grid parameters:*

    - *'max_depth': 15,*
    - *'max_features': 4,*
    - *'min_samples_leaf': 25,*
    - *'min_samples_split': 25,*
    - *'n_estimators': 200*

- *We extracted the variable importance as per Random Forests:*

| | Imp |
|---|---|
| Agency_Code | 0.323663 |
| Product Name | 0.207611 |
| Sales | 0.176532 |
| Commision | 0.112406 |
| Duration | 0.066010 |
| Type | 0.055337 |
| Age | 0.049926 |
| Destination | 0.008515 |
| Channel | 0.000000 |

SNIPPET 2. 3*: IMPORTANT VARIABLE AS PER RF*

- *Like CART, for Random Forest as well Agency code has the most importance in the model, however Sales and Product Name exchanged places.*

- *In this model, each of the variable plays a significant role in model building at some importance level and if we observe Channel Variable has the lowest importance of them all.*

### *2.2.3 NEURAL NETWORK CLASSIFIER*

- *NN is made of layers with many interconnected nodes(neurons). There are three main layers specifically;*

    - *Input Layer*
    - *Hidden Layer*
    - *Output Layer*

- *Hidden Layer can be one or more.*

- *Using the MLP Classifier and grid search and running multiple combinations of iterations, below are the best grid parameters.*

    - *'hidden_layer_sizes': 100,*
    - *'max_iter': 2500,*
    - *random_state=1*
    - *'tol': 0.01*

## 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.

### *2.3.1 CART PERFORMANCE*

- *After predicting the test and train data, below is the head of ytest_predict_prob:*

| | 0 | 1 |
|---|---|---|
| **0** | 0.617672 | 0.382328 |
| **1** | 0.690120 | 0.309880 |
| **2** | 0.538041 | 0.461959 |
| **3** | 0.237722 | 0.762278 |
| **4** | 0.817804 | 0.182196 |

SNIPPET 2. 4*: PROBABILITY OF YTEST*

- *It seems the focus is more on predicted probability of zeros(0s).*

#### *2.3.1.1 Training data*

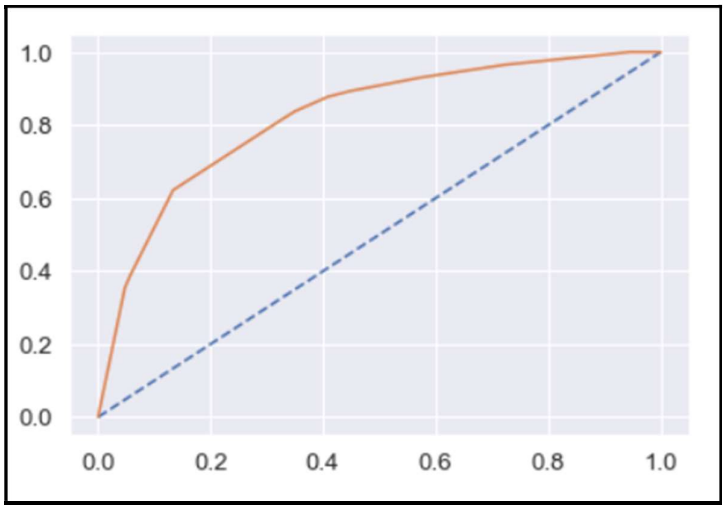


FIGURE 2. 15*: AUC AND ROC CURVE OF CART ON TRAINING DATA*

- *If we observe above figure the Area Under the Curve (AUC) is 82% which seems the model has got trained well.*

```
array([[1275,  196],
       [ 238,  391]], dtype=int64)
```

SNIPPET 2. 5*: CONFUSION MATRIX OF CART ON TRAINING DATA*

- *We can see the confusion matrix of CART on Training data that is actual vs predicted.*

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

TABLE 2. 7: *CLASSIFICATION REPORT OF CART ON TRAINING DATA*

- *From the above table we can observe the following details:*
    - *The accuracy of the training data is 79%*
    - *The Recall for the training data is 62%*
    - *The Precision for the training data is 67%*

- *Therefore, from the above results we can say that out of 100 cases 62 cases falls under true positive.*

### *2.3.1.2 Testing data*

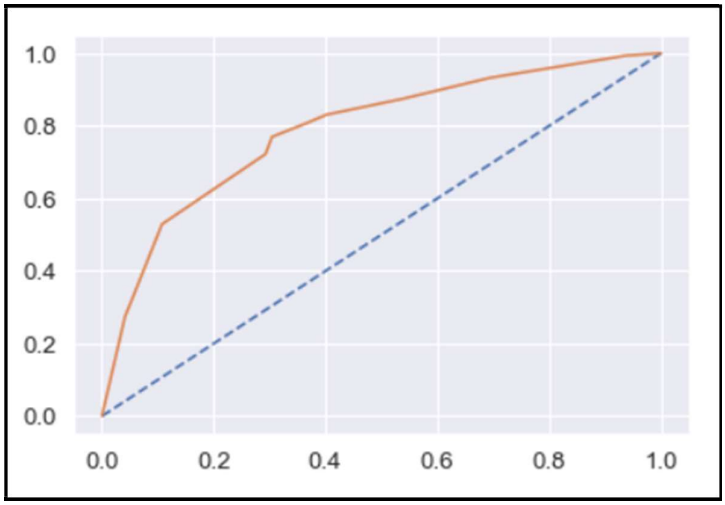


FIGURE 2. 16: *AUC AND ROC CURVE OF CART ON TESTING DATA*

- *If we observe above figure the Area Under the Curve (AUC) for the testing data is 78.6% which is above the cut of line and looks like the model had kind of performed good.*

```
array([[540,  65],
       [139, 156]], dtype=int64)
```

SNIPPET 2. 6*: CONFUSION MATRIX OF CART ON TESTING DATA*

- *We can see the confusion matrix of CART on Testing data that is actual vs predicted.*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.80 | 0.89 | 0.84 | 605 |
| 1 | 0.71 | 0.53 | 0.60 | 295 |
| accuracy | | | 0.77 | 900 |
| macro avg | 0.75 | 0.71 | 0.72 | 900 |
| weighted avg | 0.77 | 0.77 | 0.76 | 900 |

TABLE 2. 8*: CLASSIFICATION REPORT OF CART ON TESTING DATA*

- *From the above table we can observe the following details:*
  - *The accuracy of the testing data is 77%*
  - *The Recall for the testing data is 53%*
  - *The Precision for the training data is 71%*

- *Therefore, from the above results we can say that 71% of cases falls under actual true positive when the model predicts.*

### *2.3.2 RANDOM FOREST PERFORMANCE.*

#### *2.3.2.1 Training data*

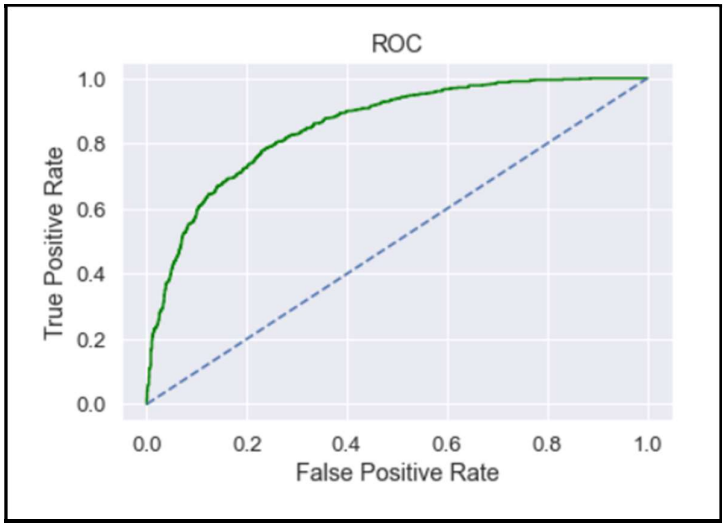


FIGURE 2. 17*: AUC AND ROC CURVE OF RF ON TRAINING DATA*

- *If we observe above figure the Area Under the Curve (AUC) is 85.4% which seems the model taken all the attributes into consideration with multiple decision tree.*

```
array([[1333,  138],
       [ 276,  353]], dtype=int64)
```

SNIPPET 2. 7*: CONFUSION MATRIX OF RF ON TRAINING DATA*

- *We can see an array of confusion matrix obtained by Random Forest on Training data that is actual vs predicted.*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.91 | 0.87 | 1471 |
| 1 | 0.72 | 0.56 | 0.63 | 629 |
| accuracy | | | 0.80 | 2100 |
| macro avg | 0.77 | 0.73 | 0.75 | 2100 |
| weighted avg | 0.80 | 0.80 | 0.80 | 2100 |

TABLE 2. 9*: CLASSIFICATION REPORT OF RF ON TRAINING DATA*

- *From the above table we can observe the following details:*

- *The accuracy of the training data is 80%*
- *The Recall for the training data is 56%*
- *The Precision for the training data is 72%*

- *Therefore, from the above results we can say the accuracy on the training data stands good and also could say that the model has evaluated the data well hence, the prediction may result to good performance.*

### *2.3.2.2 Testing data*

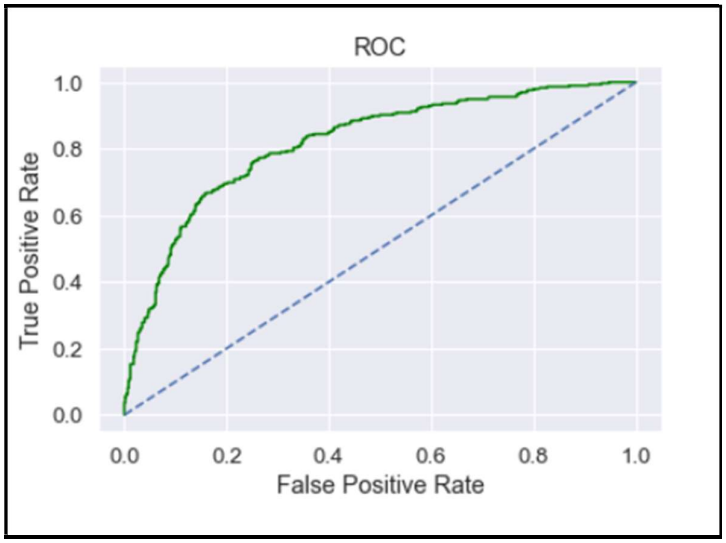


FIGURE 2. 18*: AUC AND ROC CURVE OF RF ON TESTING DATA*

- *If we observe above figure the Area Under the Curve (AUC) is 81.8% which seems the model learnt the data and the prediction seems to be not overfitted which could be considered to the level of satisfactory.*

```
array([[553,  52],
       [159, 136]], dtype=int64)
```

SNIPPET 2. 8*: CONFUSION MATRIX OF RF ON TESTING DATA*

- *We can see an array of confusion matrix obtained by Random Forest on Testing data that is actual vs predicted.*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.91 | 0.84 | 605 |
| 1 | 0.72 | 0.46 | 0.56 | 295 |
| accuracy | | | 0.77 | 900 |
| macro avg | 0.75 | 0.69 | 0.70 | 900 |
| weighted avg | 0.76 | 0.77 | 0.75 | 900 |

TABLE 2. 10*: CLASSIFICATION REPORT OF RF ON TESTING DATA*

- *From the above table we can observe the following details:*
    - *The accuracy of the testing data is 84%*
    - *The Recall for the testing data is 46%*
    - *The Precision for the testing data is 72%*

- *From the above results conclusions may be drawn that the model's performance exceptionally good as we see the accuracy rate on the testing data is upstanding also, .we could derive the model has less inadequacy hence, the prediction has resulted to a quality accomplishment.*

### *2.3.3 NEURAL NETWORK CLASSIFIER PERFORMANCE*

#### *2.3.3.1 Training data*

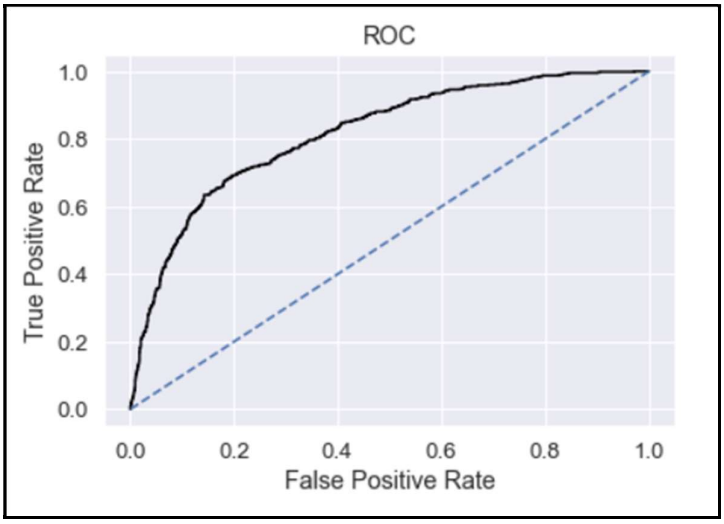


FIGURE 2. 19*: AUC AND ROC CURVE OF NN OF TRAINING DATA*

- *If we observe above figure the Area Under the Curve (AUC) is 81.5% and the model has got trained well.*

```
array([[547,  58],
       [154, 141]], dtype=int64)
```

SNIPPET 2. 9*: CONFUSION MATRIX OF NN ON TRAINING DATA*

- *We can see an array of confusion matrix obtained by ANN on Training data that is actual vs predicted.*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.83 | 0.88 | 0.85 | 1471 |
| 1 | 0.67 | 0.58 | 0.62 | 629 |
| accuracy | | | 0.79 | 2100 |
| macro avg | 0.75 | 0.73 | 0.74 | 2100 |
| weighted avg | 0.78 | 0.79 | 0.78 | 2100 |

TABLE 2. 11*: CLASSIFICATION REPORT ON TRAINING DATA*

- *From the above table we can observe the following details:*
    - *The accuracy of the training data is 79%*

- *The Recall for the training data is 58%*
- *The Precision for the training data is 67%*

- *From the above results we can say the Neural Network model has not performed on training may be cause hyperparameter was not iterated to its best fit of combination let's look forward for testing the data set.*

### *2.3.3.2 Testing data*

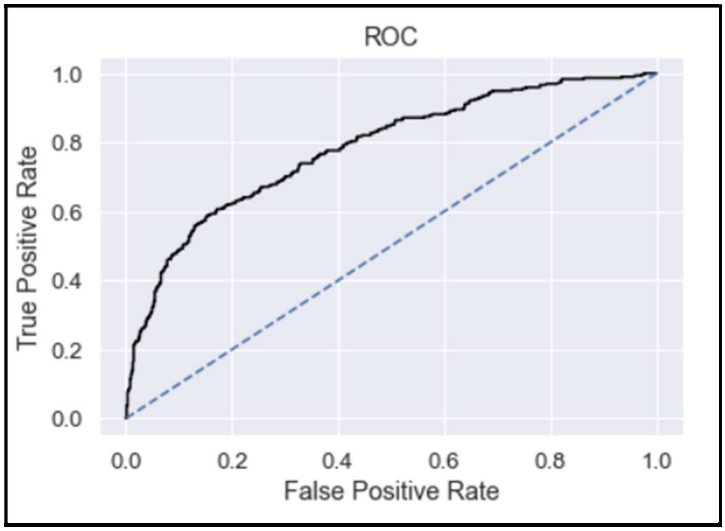


FIGURE 2. 20: *AUC AND ROC CURVE OF NN ON TESTING DATA*

- *If we observe above figure the Area Under the Curve (AUC) is 78.2% and the model has not so good when compared to the output of Random Forest model.*

```
array([[547,  58],
       [154, 141]], dtype=int64)
```

SNIPPET 2. 10: *CONFUSION MATRIX OF NN ON TESTING DATA*

- *We can see an array of confusion matrix obtained by ANN on Testing data that is actual vs predicted values.*

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.78 | 0.90 | 0.84 | 605 |
| 1 | 0.71 | 0.48 | 0.57 | 295 |
| accuracy | | | 0.76 | 900 |
| macro avg | 0.74 | 0.69 | 0.70 | 900 |
| weighted avg | 0.76 | 0.76 | 0.75 | 900 |

TABLE 2. 12: *CLASSIFICATION REPORT ON TESTING DATA*

- *From the above table we can observe the following details:*
  - *The accuracy on the testing data is 76%*
  - *The Recall for the testing data is 48%*
  - *The Precision for the testing data is 71%*

- *Therefore, from the above results we can say the Neural Network model has not performed well when compared to the Random Forest model.*

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

- *Below we are comparing Accuracy, AUC, Recall, Precision and F1 score of all the models, where Target is 0, i.e., the claimed as NO.*

- *The logic to choose Claimed as NO is that the model is calculating Claimed as No more accurately than Claimed as Yes. Also, this way we will be able to identify using the attributes that which policy will not be claimed with more than approx. 75% accuracy.*

| | CART Train | CART Test | Random Forest Train | Random Forest Test | Neural Network Train | Neural Network Test |
|---|---|---|---|---|---|---|
| **Accuracy** | 0.79 | 0.77 | 0.80 | 0.77 | 0.79 | 0.76 |
| **AUC** | 0.82 | 0.79 | 0.85 | 0.82 | 0.82 | 0.78 |
| **Recall** | 0.62 | 0.53 | 0.56 | 0.46 | 0.58 | 0.48 |
| **Precision** | 0.67 | 0.71 | 0.72 | 0.72 | 0.67 | 0.71 |
| **F1 Score** | 0.64 | 0.60 | 0.63 | 0.56 | 0.62 | 0.57 |

TABLE 2. 13: *COMPARISON OF ALL 3 MODELS*

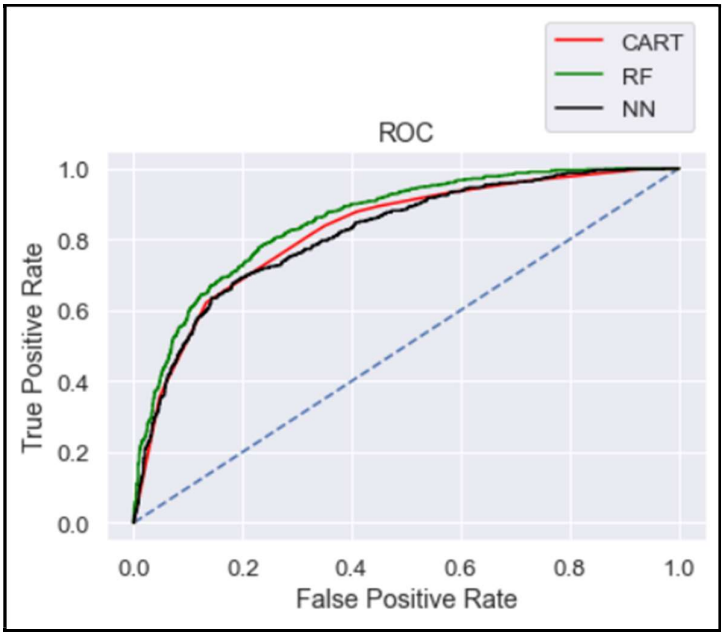


FIGURE 2. 21*: ROC CURVES OF TRAIN DATA ON ALL THE THREE MODELS*

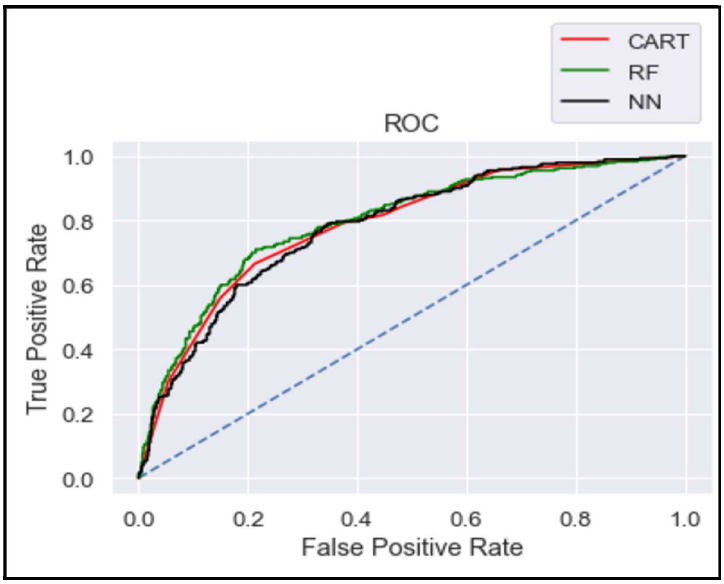


FIGURE 2. 22*: ROC CURVES OF TEST DATA ON ALL THE THREE MODELS*

- *Out of the 3 models, Random Forest has slightly better performance than the Cart and Neural network model.*

- *Overall, all the 3 models are reasonably stable enough to be used for making any future predictions. From Random Forest Model, the variable change is found to be the most useful feature amongst all other features*

*for predicting if a person will claim or not. If change is NO, then those policies have more chances of getting claimed.*

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

- *I strongly recommended we collect more real time data if possible. So, we may dive deep into the process to understand the workflow and could predict more precisely.*

- *It seems that all the models show high accuracy in predicting the customers who will not claim for tour insurance after performing the supervised learning algorithm.*

- *As we know this Insurance firm was facing higher claim frequency and this model would certainly help in reducing the ratio when compared to before and after.*

- *Since the variable Agency code seem to be the most important factor in deriving the model, therefore I recommend that the insurance company tie up with more Agencies to expand its business.*

- *To attain the less frequency of claims, they should add certain steps to their policy's terms and conditions that would benefit both customers and company.*

- *Using this model and customer data this insurance firm can easily pick their profitable customers and target could be properly decided.*

- *Then business can easily target the customers who will not claim for tour insurance. Once Team receives customer data who falls under NO claim status as per the model, then team needs to build strong relationship with those customers because you only get profit when repeated customer sees loyalty and trust in an organization.*

# Project Data Mining

- *I believe that the tour insurance company should also increase its varieties for Product Name. For now, they are having Bronze, Cancellation, Customized, Gold and Silver plans, but adding few more to the list will encourage customers to choose the optimum plan which proves to be the successful for them and in return, would lead to less frequency of claims for the company. The same would also result in more sales for the tour insurance company.*

- *Product plan which has higher commission rate can be recommended to the set customers who will fall under NO claim status.*

- *Also based on the model we are getting 80% accuracy, so we need customer books airline tickets or plans, cross sell the insurance based on the claim data pattern. Other interesting fact is more sales happen via Agency than Airlines and the trend shows the claim are processed more at Airline.*