Business Report

SMDM Project Business Report DSBA

***Sanjay Srinivasan***

***PGP-DSBA Online***

***JULY’ 21 Batch***

***Date: 21-11-2021***

# *INDEX*

|  |  |  |
| --- | --- | --- |
| ***S. No*** | ***Contents*** | ***Page* No** |
| **1.** | **Problem - 1** | 5 |
|  | Summary | 5 |
|  | Introduction | 5 |
|  | Data Description | 5 |
|  | Sample Dataset | 5 |
|  | Exploratory Data Analysis | 5 |
|  | Checking for missing values in the dataset | 6 |
|  | 1) Read the data, do the necessary initial steps, and exploratory data analysis (Uni-variate, Bi-variate, and multivariate analysis). | 6 |
|  | 2) Do you think scaling is necessary for clustering in this case? Justify | 8 |
|  | 3) Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them. | 9 |
|  | 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 | 10 |
|  | 5) Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters. | 12 |
|  |  |  |
| **2.** | **Problem - 2** | 14 |
|  | Summary | 14 |
|  | Introduction | 14 |
|  | Data Description | 14 |
|  | Sample Dataset | 14 |
|  | Exploratory Data Analysis | 15 |
|  | Checking for missing values in the dataset | 15 |
|  | 1) Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis). | 15 |
|  | 2) Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network. | 22 |
|  | 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. | 23 |
|  | 4) Final Model: Compare all the models and write an inference which model is best/optimized. | 29 |
|  | 5)Inference: Based on the whole Analysis, what are the business insights and recommendations | 29 |

# *List Of Tables*

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Dataset sample | 5 |
| 1.2 | Datatypes of the variable | 5 |
| 1.3 | Check null values | 6 |
| 1.4 | Sample dataframe after linkage in hierarchical clustering | 10 |
| 1.5 | Grouping linkage based on hierarchical clustering | 10 |
| 1.6 | linkage based on K - means clustering | 11 |
| 1.7 | Grouping linkage based on K - means clustering | 12 |
| 1.8 | Count k – means linkage into dataframe | 12 |
| 1.9 | Grouping k – means linkage into dataframe | 12 |
| 2.1 | Dataset Sample | 14 |
| 2.2 | Datatypes of the variable | 15 |
| 2.3 | Check null values | 15 |
| 2.4 | Check data types of the dataframe after converting into integer. | 22 |
| 2.5 | Dataframe for train dataset | 22 |
| 2.6 | Standardising the dataframe | 22 |

***List Of Figures***

|  |  |  |
| --- | --- | --- |
| ***S.No*** | ***Content*** | ***Page No*** |
| 1.1 | Univariate Analysis | 6 |
| 1.2 | Bivariate Analysis | 7 |
| 1.3 | Multivariate Analysis for spending’s vs. credit limit vs. current balance | 7 |
| 1.4 | Multivariate analysis of pairplot | 8 |
| 1.5 | Before Scaling | 8 |
| 1.6 | After Scaling | 9 |
| 1.7 | Before pruning | 9 |
| 1.8 | After pruning | 9 |
| 1.9 | Pruning to the last 10 values | 10 |
| 1.10 | K – means inertia value | 10 |
| 1.11 | Elbow curve for K - means | 11 |
| 1.12 | Silhouette score | 11 |
| 1.13 | Silhouette width | 11 |
| 1.14 | Silhouette Sample value | 12 |
| 1.15 | Count k – means linkage | 12 |
| 2.1 | Categorical value count | 16 |
| 2.2 | Number of duplicate rows and sample duplicate rows | 16 |
| 2.3 | Dataframe before treating outliers | 17 |
| 2.4 | After treating outliers | 18 |
| 2.5 | Univariate analysis for Type | 18 |
| 2.6 | Univariate analysis for Claimed | 19 |
| 2.7 | Bivariate analysis Type vs. Sales | 19 |
| 2.8 | Bivariate analysis Destination vs. Sales | 20 |
| 2.9 | Multivariate analysis of pairplot | 20 |
| 2.10 | Multivariate analysis heatmap | 21 |
| 2.11 | Categorical values to categorical codes | 21 |
| 2.12 | Shape of trained and test dataframe | 22 |
| 2.13 | Parameters for GridsearchCV in Decision Tree Classifier | 23 |
| 2.14 | Best parameter for Decision tree classifier | 23 |
| 2.15 | Parameters for GridsearchCV in Random Forest Algorithm | 23 |
| 2.16 | Best parameter for Random Forest Algorithm | 23 |
| 2.17 | Parameters for GridsearchCV in Artificial Neural Networks Algorithm | 23 |
| 2.18 | Best parameter for Artificial Neural Networks Algorithm | 23 |
| 2.19 | Predicted values from the train dataset of CART model | 24 |
| 2.20 | confusion matrix from Train data of CART model | 24 |
| 2.21 | confusion matrix from test data of CART model | 24 |
| 2.22 | Classification Report from train data of CART model | 24 |
| 2.23 | Classification Report from test data of CART model | 24 |
| 2.24 | AUC and ROC curve train data of CART model | 25 |
| 2.25 | AUC and ROC curve test data of CART model | 25 |
| 2.26 | Predicted values from the train dataset of RF model | 26 |
| 2.27 | confusion matrix from Train data of RF Model | 26 |
| 2.28 | confusion matrix from Test data of RF Model | 26 |
| 2.29 | Classification Report from train data of RF Model | 26 |
| 2.30 | Classification Report from test data of RF Model | 26 |
| 2.31 | AUC and ROC curve train data of RF model | 27 |
| 2.32 | AUC and ROC curve test data of RF model | 27 |
| 2.33 | Predicted values from the train dataset of ANN model | 27 |
| 2.34 | confusion matrix from Train data of ANN Model | 28 |
| 2.35 | confusion matrix from Test data of ANN Model | 28 |
| 2.36 | Classification Report from train data of ANN model | 28 |
| 2.37 | Classification Report from test data of ANN model | 28 |
| 2.38 | AUC and ROC curve train data of ANN model | 28 |
| 2.39 | AUC and ROC curve train data of ANN model | 29 |

# Problem - 1

*Summary*

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage of 210 customers. From this 210 customers data, we group them based on the spending’s , current balance, amount paid in advance

*Introduction*

The purpose of this exercise is to explore the dataset and make the clustering for these customers, based on the group we can categorize the promotional offer to increase their usage of credit cards.

*Data Description*

* Spending: Amount spent by the customer per month (in 1000s)
* Advance payments: Amount paid by the customer in advance by cash (in 100s)
* Probability of full payment: Probability of payment done in full by the customer to the bank
* Current balance: Balance amount left in the account to make purchases (in 1000s)
* Credit limit: Limit of the amount in credit card (10000s)
* Min payment amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
* Max spent in single shopping: Maximum amount spent in one purchase (in 1000s)

*Sample of the dataset:*

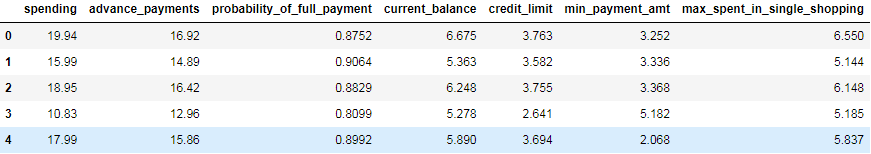


Table 1.1 Dataset Sample

*Exploratory Data Analysis*

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

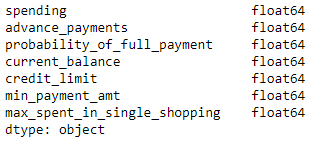


Table- 1.2. Datatypes of the variable

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

*Check for missing values in the dataset:*

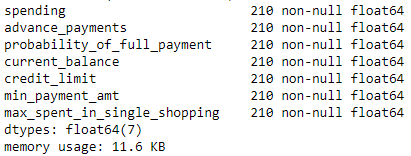


Table- 1.3. Check null values

### Problem 1A:

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

***Uni-Variate Analysis:***

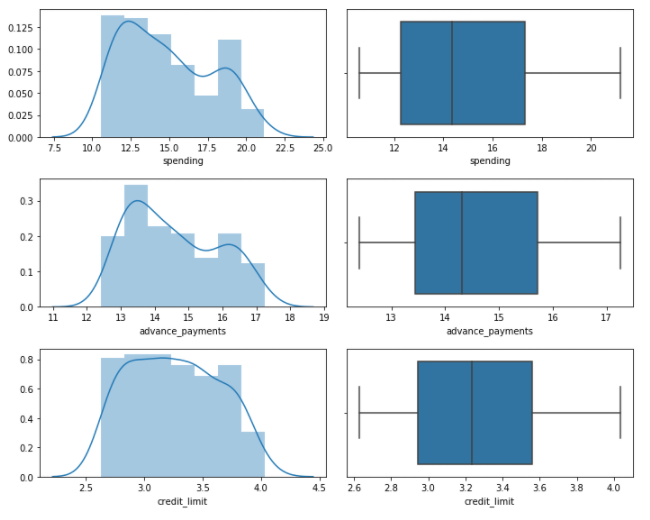


Fig – 1.1 Univariate Analysis

From the above chart (displot and boxplot), there are no outliers.

***Bi – variate Analysis:***

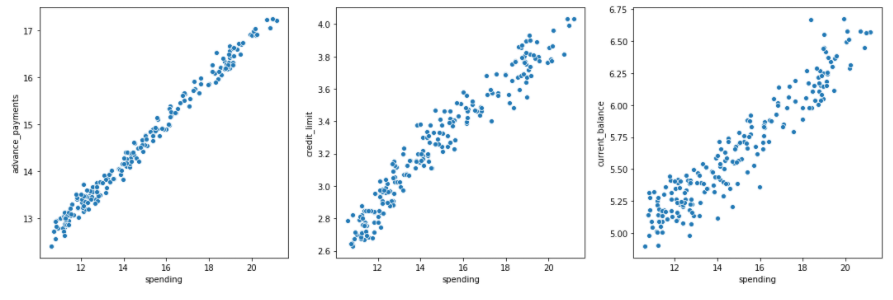


Fig – 1.2 Bivariate Analysis

### From the scatterplot, we can infer that as the 'spending' increases, the 'advance\_payments', 'credit\_limit', 'current\_balance' increases. From this, the customer has the high 'current\_balance' have the high 'spending'.

***Multi – variate Analysis:***

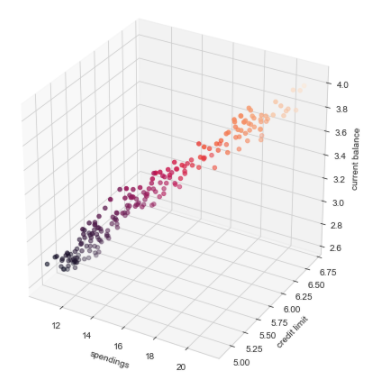


Fig – 1.3 Multivariate Analysis for spending’s vs. credit limit vs. current balance

From this 3D Scatter plot, we can infer that, If the spending increases, the credit limit and current balance also increases.

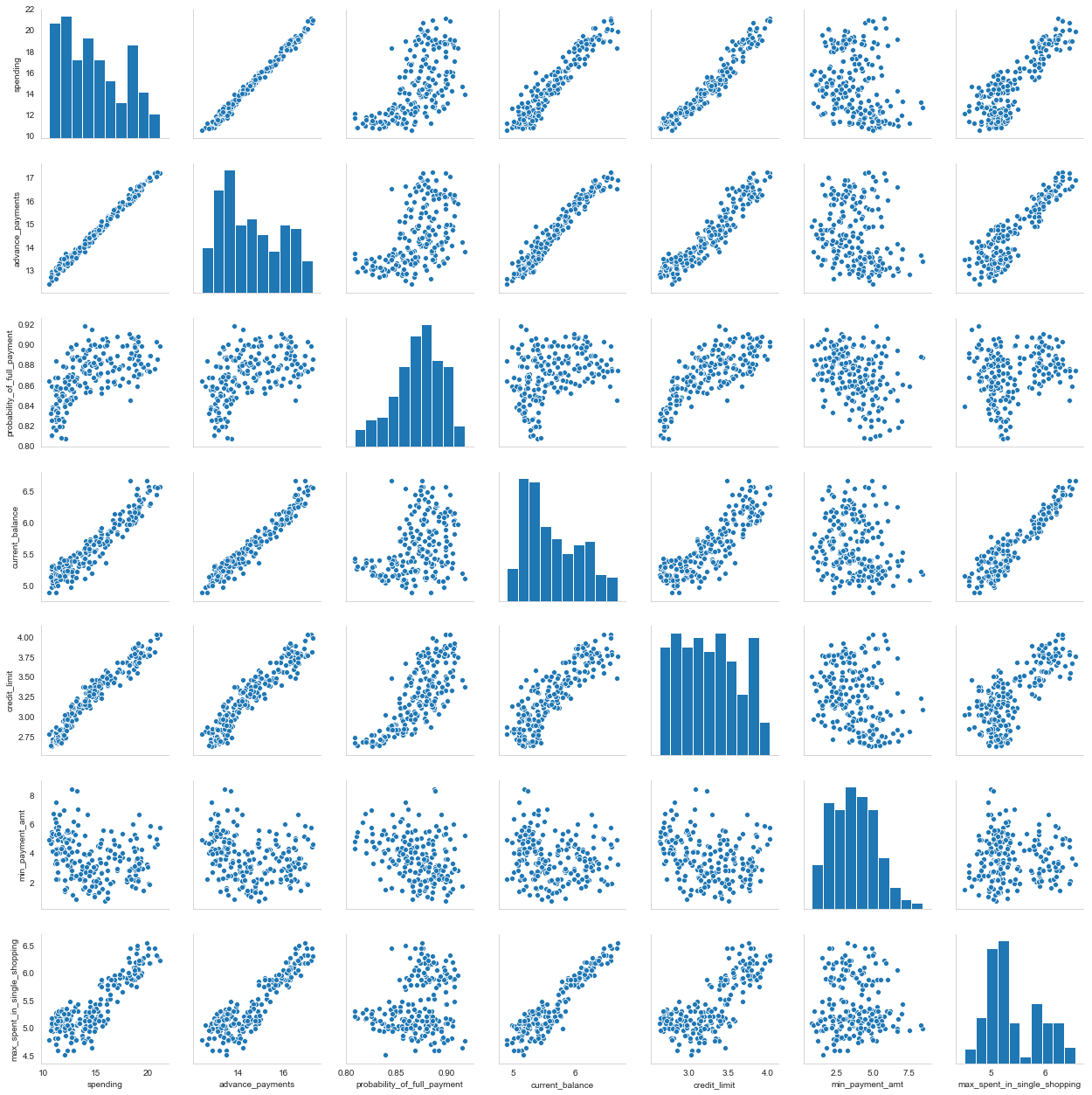
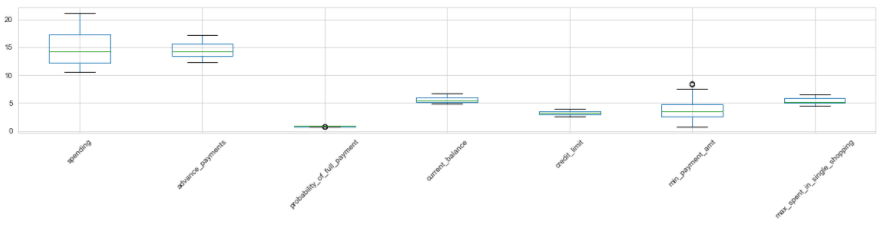


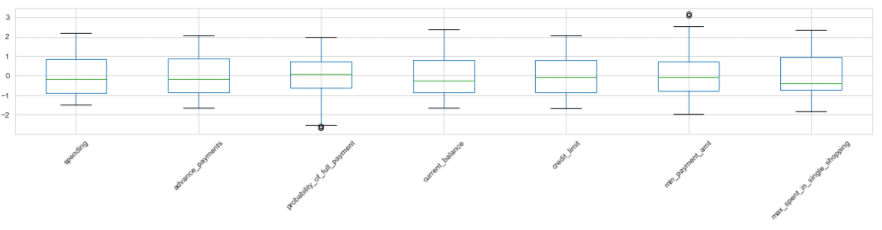
Fig – 1.4 Multivariate analysis of pairplot

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

***Before Scaling:***

 Fig – 1.5 Before Scaling

***After Scaling:***

 Fig – 1.6 After Scaling

Yes, Scaling needs to be done as the values of the variables are different. spending, advance\_payments are in different values and this may get more weightage. The plot of the data prior and after scaling. Scaling will have all the values in the relative same range. I have used z-score to standarised the data to relative same scale -3 to +3

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

***Before pruning:***

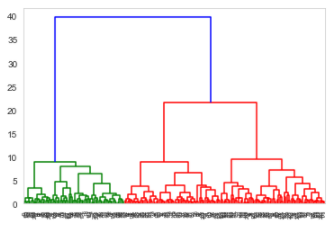


Fig – 1.7 Before pruning

From the above dendogram, we can infer that there are two clusters are present in the data.

***After pruning:***

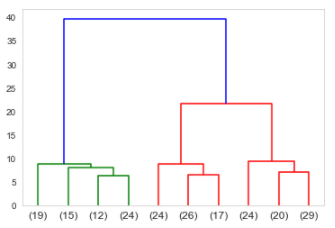


Fig – 1.8 After pruning

We are pruning the last 10 values of the dendogram.

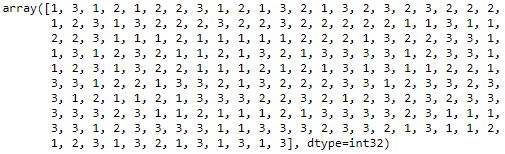


Fig – 1.9 Pruning to the last 10 values

The cluster grouping linkage based on the dendrogram, 3 or 4 looks good. The further analysis, and based on the dataset had gone for 3 linkage solution based on the hierarchical clustering.

Sample data of linkage based on hierarchical clustering.

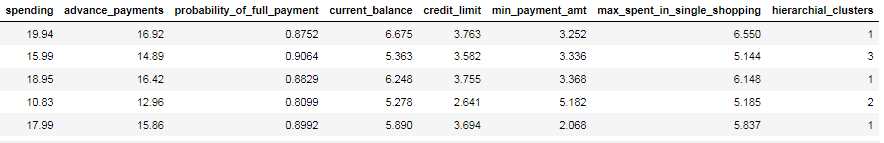


Table – 1.4 Sample dataframe after linkage in hierarchical clustering

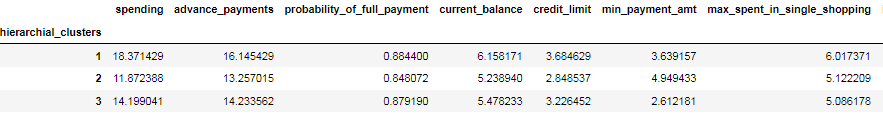


Table – 1.5 Grouping linkage based on hierarchical clustering

From this hierarchical clustering, we can infer that high Spending has been done with medium frequency. Moderate spending of customer are high with frequency when compared with other clusters.

### 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.

In K-means clustering, the inertia is calculated, from that finding plotting the elbow curve and finding the clusters from the elbow curve

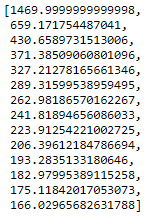


Fig – 1.10 K – means inertia value

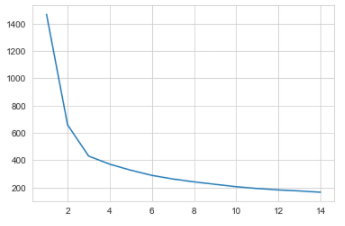


Fig – 1.11 Elbow curve for K - means

The k-means cluster linkage value is concatenated in dataframe. Sample dataframe is shown below.

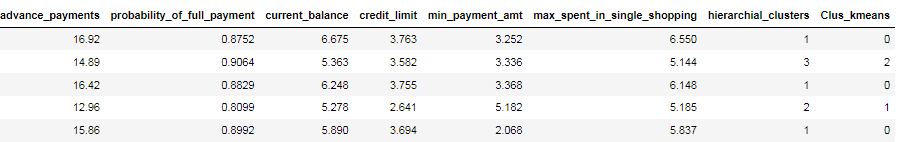


Table – 1.6 linkage based on K - means clustering

***Silhouette Score and Silhouette width:***



Fig –1.12 Silhouette score

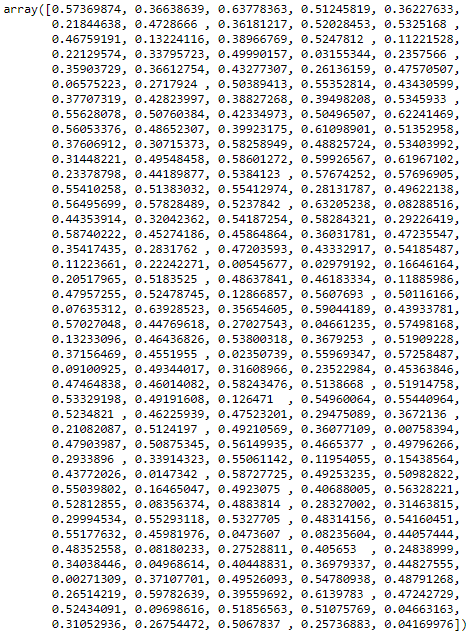


Fig – 1.13 Silhouette width

***Silhouette sample value:***



Fig – 1.14 Silhouette Sample value

From the elbow curve and the silhouette sample output 3 is the ideal clusters for the data

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

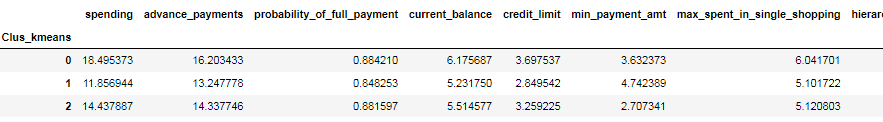


Table – 1.7Grouping linkage based on K - means clustering

Calculating the number of K-means cluster percentage and loading into the dataframe.

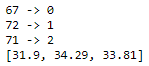


Fig – 1.15 Count k – means linkage

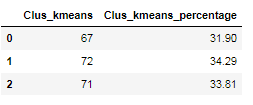


Table- 1.8. Count k – means linkage into dataframe

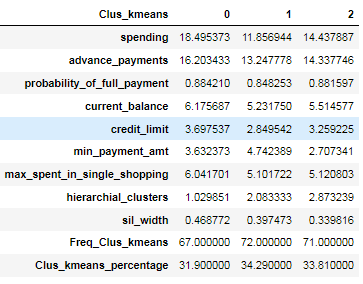


Table- 1.9. Grouping k – means linkage into dataframe

Inference from the above, Group 2 has the highest spending, whereas Group 1 has the least spending.

***Promotional offer High spending group:***

* Giving any reward points might increase their purchases.
* Maximum ”**max\_spent\_in\_single\_shopping** “ is high for this group, so can be offered discount/offer on next transactions upon full payment
* Increase their credit limit
* Increase spending habits
* Give loan against the credit card, as they are customers with good repayment record.
* Tie up with luxury brands, which will drive more “**one\_time\_maximum**” spending

###### **Promotional offer Medium spending group:**

* They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate.
* Increase spending habits by trying with premium Travel tickets, luxurious stay as this will encourage them to spend more

###### **Promotional offer Low spending group:**

* Customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.

# Problem – 2

*Summary*

The data is gathered from an Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. The task is to make a model which predicts the claim status and provide recommendations to management. By using CART, RF & ANN model and comparing the models' performances by training and testing the dataset.

*Introduction*

The purpose of this exercise to create model using CART (Decision Tree), RF (Random Forest), ANN (Artificial Neural Networks) algorithm and predict the claim status and provide recommendations from the dataset. This dataset consist of 3000 rows and 10 columns,

*Data Description*

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

*Sample of the dataset:*

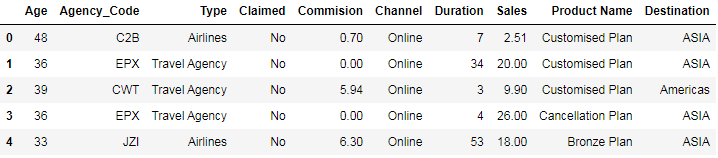


Table 2.1 Dataset Sample

Dataset has 10 variables with insurance details. Based on the insurance details, who have enrolled in different colleges is defined.

*Exploratory Data Analysis*

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

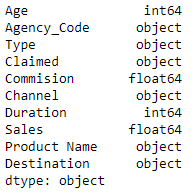


Table 2.2 Datatypes of the variable

There are total 3000 rows and 10 columns in the dataset. Out of 10, 2 column is of integer type, 2 column is of float (Decimal value) type and rest 6 are of integer data type.

*Check for missing values in the dataset:*

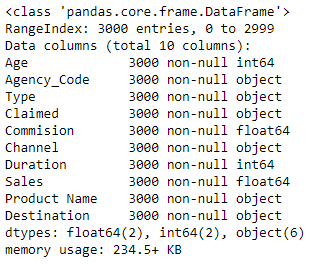


Table 2.3 Check null values

From this, it is clear that there are no null values present in the dataset.

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

The numbers of unique variables are taken from the categorical column.

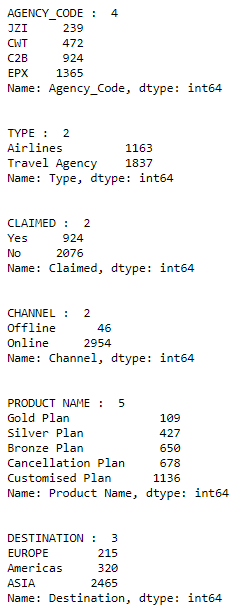


Fig – 2.1 Categorical value count

The numbers of duplicate values are taken from the dataset and duplicate records have been dropped from the dataset.

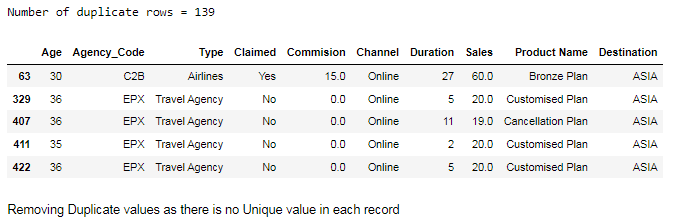


Fig – 2.2 Number of duplicate rows and sample duplicate rows

Finding the outliers for the Numerical data from the dataset

***Before Treating Outliers:***

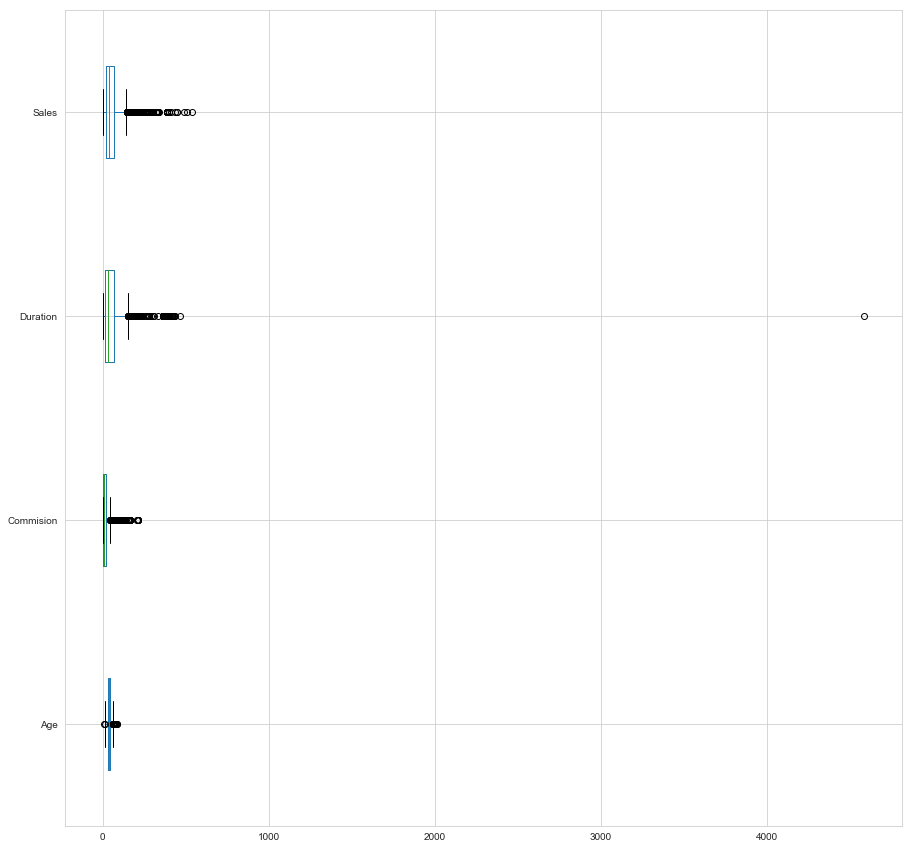


Fig – 2.3 Dataframe before treating outliers

***After Treating Outliers:***

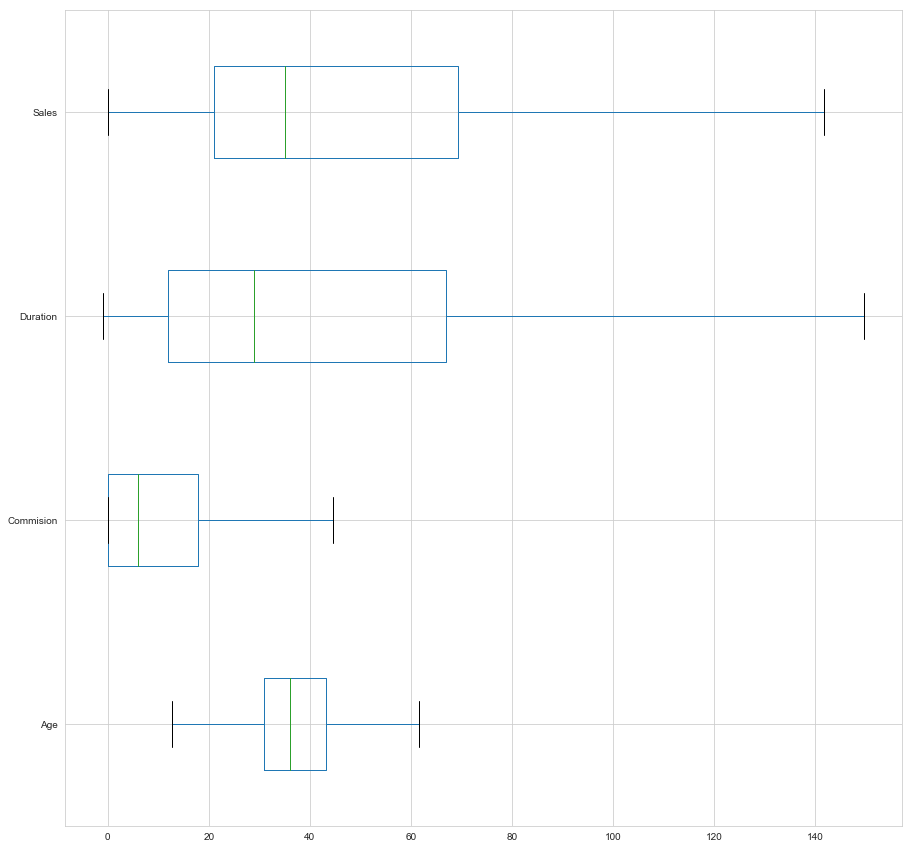
******

Fig – 2.4 After treating outliers

#### Univariate Analysis:

Univariate analysis is the simplest form of analysing data. Analysing each variable in detail.

Insights:

1163 customers have opted for Airlines and 1867 customers have opted for Airlines and Travel Agency

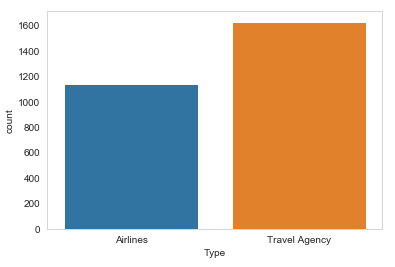


Fig – 2.5 Univariate analysis for Type

The value counts from the claimed variable are, 924 customers have claimed the insurance whereas 2076 customers have not claimed the insurance.

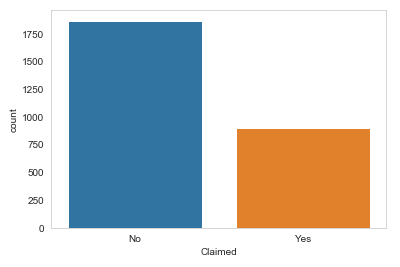


Fig – 2.6 Univariate analysis for Claimed

***Bivariate Analysis:***

Bivariate analysis is the simplest form of analysing data. Analysing a single variable with another variable in detail.

*Insights:*

The Amount of sales per customer for procuring travel insurance with the type of insurance that customer has purchased. From the below boxplot, we can infer that customers purchased through Airlines bought more travel insurance than the customers bought the travel insurance from the Travel agency.

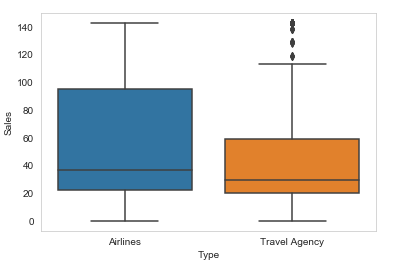


Fig – 2.7 Bivariate analysis Type vs. Sales

The customers traveling to their destination with the type of insurance that customer has purchased. From the below boxplot, we can infer that customers have purchased the travel insurance for reaching their destination. The more travel insurance is bought by the customers travelling to the America as their destination.

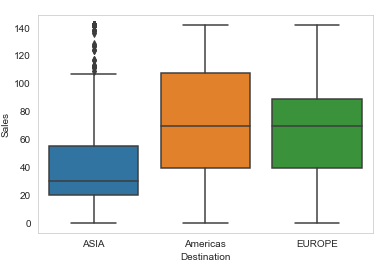


Fig – 2.8 Bivariate analysis Destination vs. Sales

**Multivariate Analysis:**

Analysing the data with two or more variable.

*Insights:*

There is a strong correlation observed between few fields.

From the pairplot and correlation, we can infer that, the Strong correlation is between the sales and commission.

And weak correlation is between the Age and Duration

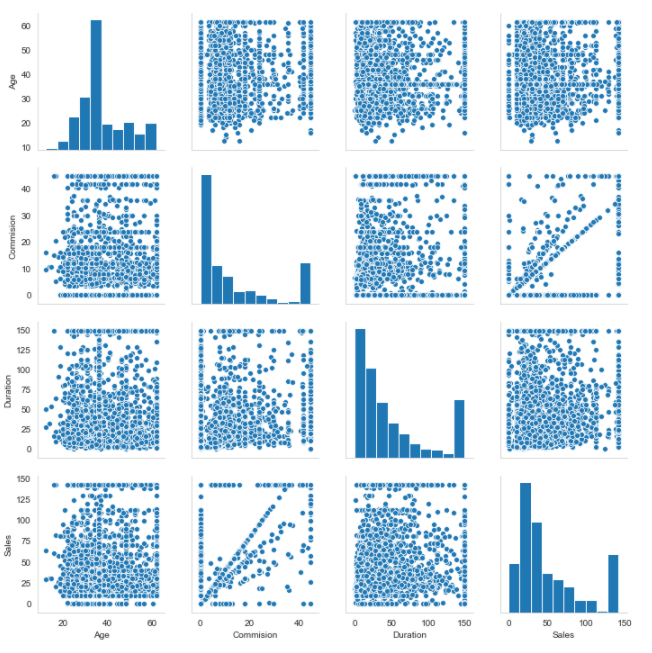


Fig – 2.9 Multivariate analysis of pairplot



Fig – 2.10 Multivariate analysis heatmap

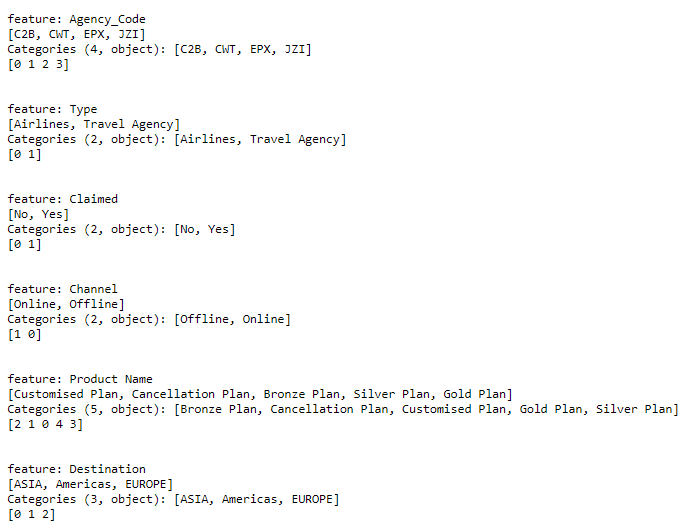
The Categorical variable is converted into the numerical variable (Integer) 

Fig – 2.11 Categorical values to categorical codes

The dataframe is initialized after dropping the duplicate records, treating the outliers and converting the categorical value into integer value.

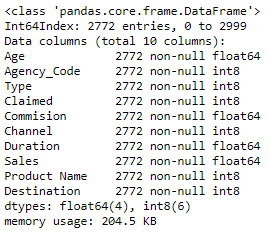


Table – 2.4 Check data types of the dataframe after converting into integer.

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

The sample dataframe is shown below, after dropping the target variable from the dataframe and before scaling is performed.

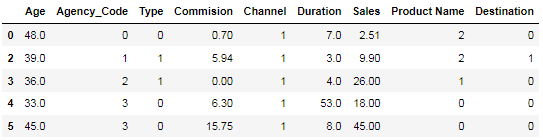


Table 2.5 Dataframe for train dataset

The data in the dataframe has been standardised to the same range using the standardScalar method.

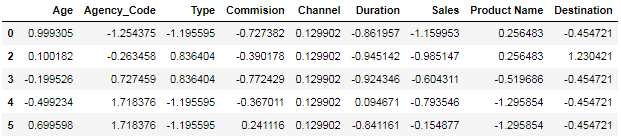


Table 2.6 Standardising the dataframe

The dataframe are split into train and test data from the dataframe. The train data has 70% of the data and test data has 30% of the data from the dataframe.

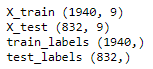


Fig – 2.12 Shape of trained and test dataframe

***Decision Tree Classifier:***

The decision tree classifier with the parameters using the grid search CV. The model is fitted into the decision tree algorithm using the grid search CV.

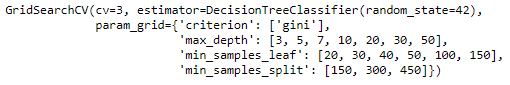


Fig – 2.13 Parameters for GridsearchCV in Decision Tree Classifier

The best parameters are identified from the decision tree algorithm by using the grid search CV.

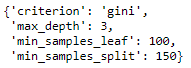


Fig – 2.14 Best parameter for Decision tree classifier

***Random Forest Algorithm:***

The Random forest algorithm with the parameters using the grid search CV. The model is fitted into the Random forest algorithm using the grid search CV.

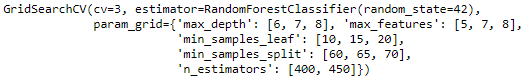


Fig – 2.15 Parameters for GridsearchCV in Random Forest Algorithm

The best parameters are identified from the Random forest algorithm by using the grid search CV.

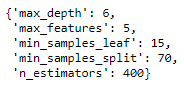


Fig – 2.16 Best parameter for Random Forest Algorithm

***Artificial Neural Networks Algorithm:***

The Artificial Neural Networks algorithm with the parameters using the grid search CV. The model is fitted into the Artificial Neural Networks algorithm using the grid search CV.

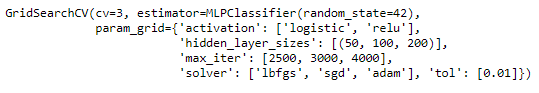


Fig – 2.17 Parameters for GridsearchCV in Artificial Neural Networks Algorithm

The best parameters are identified from the Artificial Neural Networks algorithm by using the grid search CV.

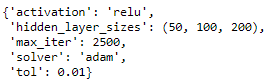


Fig – 2.18 Best parameter for Artificial Neural Networks Algorithm

**2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.**

***Decision Tree Classifier:***

The values are predicted from the train data.

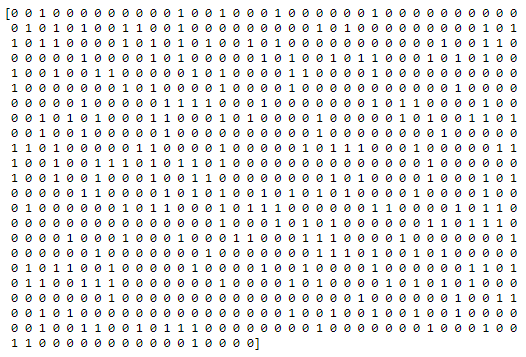


Fig – 2.19 Predicted values from the train dataset of CART model

Confusion Matrix is obtained from the train data and test data using decision tree classifier.



Fig 2.20 confusion matrix from Train data of CART model



Fig 2.21 confusion matrix from test data of CART model

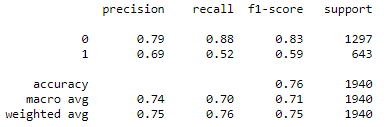


Fig 2.22 Classification Report from train data of CART model

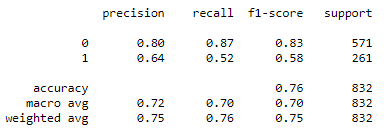


Fig 2.23 Classification Report from test data of CART model

**ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

* True Positive Rate
* False Positive Rate

The probability of the Area under the ROC curve for the train data is 79.4%

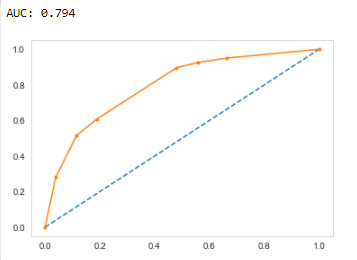


Fig 2.24 AUC and ROC curve train data of CART model

The probability of the Area under the ROC curve for the train data is 79.4%

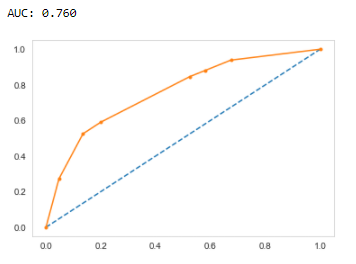


Fig 2.25 AUC and ROC curve test data of CART model

***Random Forest Algorithm:***

The values are predicted from the train data.

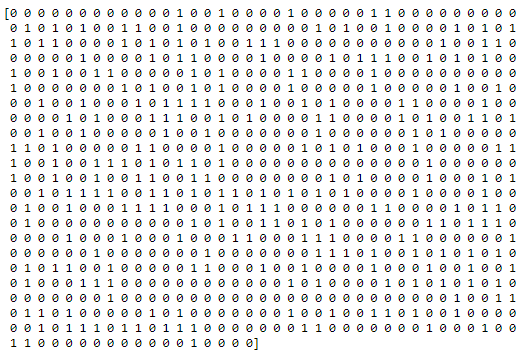


Fig – 2.26 Predicted values from the train dataset of RF model

Confusion Matrix is obtained from the train data and test data using Random Forest Algorithm.



Fig 2.27 confusion matrix from Train data of RF Model



Fig 2.28 confusion matrix from test data of RF Model

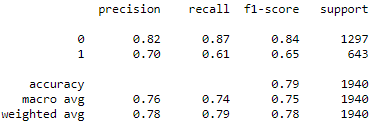


Fig 2.29 Classification Report from train data of RF Model

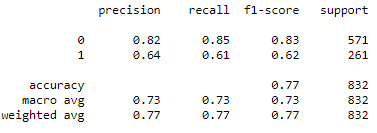


Fig 2.30 Classification Report from test data of RF Model

**ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

* True Positive Rate
* False Positive Rate

The probability of the Area under the ROC curve for the train data is 84.4%

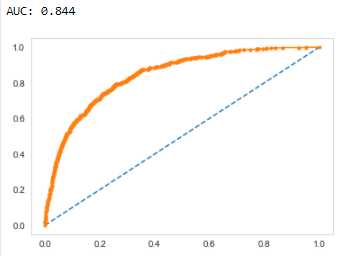


Fig 2.31 AUC and ROC curve train data of RF model

The probability of the Area under the ROC curve for the train data is 78.3%

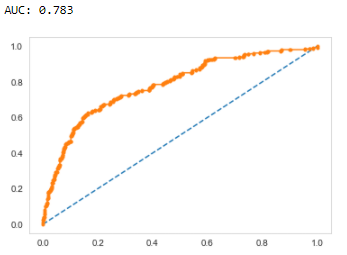


Fig 2.32 AUC and ROC curve test data of RF model

***Artificial Neural Network Algorithm:***

The values are predicted from the train data.

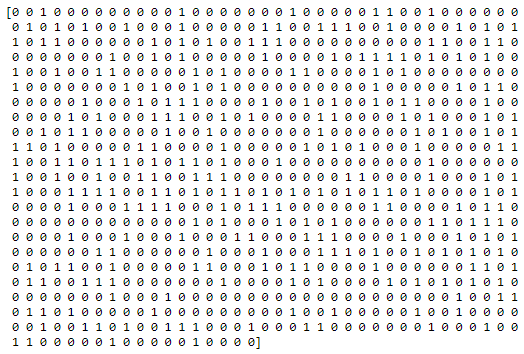


Fig – 2.33 Predicted values from the train dataset of ANN model

Confusion Matrix is obtained from the train data and test data using Artificial Neural Network Algorithm.



Fig 2.34 confusion matrix from Train data of ANN Model



Fig 2.35 confusion matrix from Test data of ANN Model

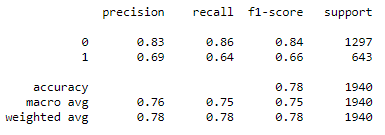


Fig 2.36 Classification Report from train data of ANN model

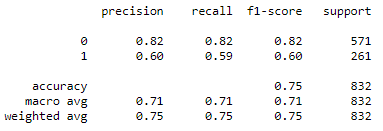


Fig 2.37 Classification Report from test data of ANN model

**ROC curve** (**receiver operating characteristic curve**) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

* True Positive Rate
* False Positive Rate

The probability of the Area under the ROC curve for the train data is 83.5%

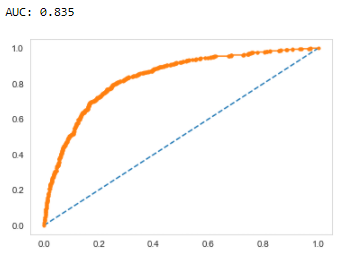


Fig 2.38 AUC and ROC curve train data of ANN model

The probability of the Area under the ROC curve for the train data is 78.3%

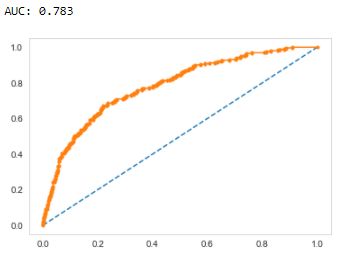


Fig 2.39 AUC and ROC curve train data of ANN model

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

Inference from the above Confusion matrix, classification report, Auc and Roc Curve, we can conclude that, for Random Forest classifier has the Highest accuracy in both train and test data.

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

***Business Insights And Recommendations:***

I strongly recommended we collect more real time unstructured data and past data if possible.

This is understood by looking at the insurance data by drawing relations between different variables such as day of the incident, time and associating it with other external information such as weather information, airline/vehicle types, etc.

* Streamlining online experiences benefitted customers, leading to an increase in conversions, which subsequently raised profits.
* As per the data 90% of insurance is done by online channel.

Key performance indicators (KPI) The KPI’s of insurance claims are:

* Reduce claims cycle time
* Increase customer satisfaction
* Combat fraud