5/21/2022

Data Mining Business Report

Sharjil Shah

PGPDSBA – JaN’21, Greatlearning

Contents

[Problem 1: Clustering 3](#_Toc104003773)

[Problem 2: CART-RF-ANN 3](#_Toc104003774)

[SOLUTIONS 4](#_Toc104003775)

[1.1 To perform Exploratory Data Analysis on the dataset and describe it briefly. 4](#_Toc104003776)

[**Importing the Dataset** 4](#_Toc104003777)

[**Dimension of the Dataset** 5](#_Toc104003778)

[**Summary of the Dataset** 5](#_Toc104003779)

[**Univariate Analysis and Bivariate Analysis** 6](#_Toc104003780)

[**Pair Plot for all the variables** 7](#_Toc104003781)

[1.2 To provide justification whether scaling is necessary for clustering in this case 7](#_Toc104003782)

[1.3 To perform hierarchical clustering to scaled data and identify the number of optimum clusters using Dendrogram and briefly describe them. 8](#_Toc104003783)

[1.4 To perform K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. 10](#_Toc104003784)

[1.5. To describe cluster profiles for the clusters defined and recommend different promotional strategies for different clusters. 11](#_Toc104003785)

[SOLUTIONS 13](#_Toc104003786)

[2.1 To read the dataset and perform the descriptive statistics and do null value condition check and write an inference on it. 13](#_Toc104003787)

[**Summary of the Dataset** 14](#_Toc104003788)

[**Checking for Missing Values** 14](#_Toc104003789)

[2.2. To split the data into test and train, build classification model CART, Random Forest and Artificial Neural Network. 18](#_Toc104003790)

[18](#_Toc104003791)

[CART Model 19](#_Toc104003792)

[Random Forest 20](#_Toc104003793)

[Artificial Neural Network (ANN) 20](#_Toc104003794)

[2.3 To check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model. 20](#_Toc104003795)

[**OC\_AUC Score and ROC Curve** 21](#_Toc104003796)

[2.4 To compare all the models and write an inference which model is best/optimized. 24](#_Toc104003797)

[**Insights:** 24](#_Toc104003798)

[2.5 To provide business insights and recommendations. 24](#_Toc104003799)

**PROJECT OBJECTIVE**

## Problem 1: Clustering

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

* 1. To perform Exploratory Data Analysis on the dataset and describe it briefly.
  2. To provide justification whether scaling is necessary for clustering in this case.
  3. To perform hierarchical clustering to scaled data and identify the number of optimum clusters using Dendrogram and briefly describe them.
  4. To perform K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.
  5. To describe cluster profiles for the clusters defined and recommend different promotional strategies for different clusters.

## 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.

* 1. To read the dataset and perform the descriptive statistics and do null value condition check and write an inference on it.
  2. To split the data into test and train, build classification model CART, Random Forest and Artificial Neural Network.
  3. To check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model.
  4. To compare all the models and write an inference which model is best/optimized.
  5. To provide business insights and recommendations.

**PROBLEM 1: CLUSTERING**

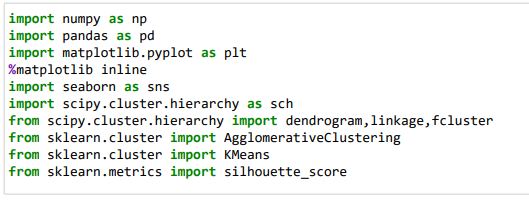
**ASSUMPTIONS**

The dataset provided to us is stored as “bank\_marketing\_part1\_Data.csv” which contains data of 210 customers and 7 variables namely:

|  |  |
| --- | --- |
| 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) |

**IMPORTING PACKAGES**

So as to import the dataset and perform Exploratory Data Analysis on the given dataset we imported the following packages:

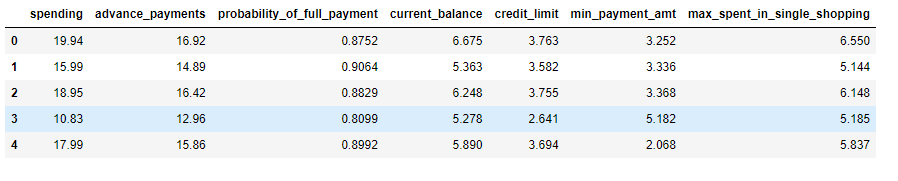


# SOLUTIONS

## To perform Exploratory Data Analysis on the dataset and describe it briefly.

### **Importing the Dataset**

The dataset in question is imported in jupyter notebook and will store the dataset in “**bank\_df**”. The top 5 rows of the dataset are viewed.



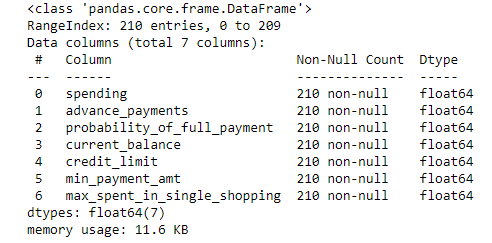
### **Dimension of the Dataset**

The dataset has 210 columns and 7 rows



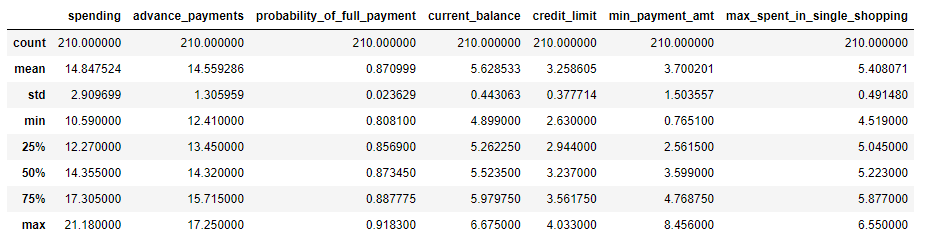
**Structure of the Dataset**

Structure of the Dataset is computed using the info function and below is the output is observed.



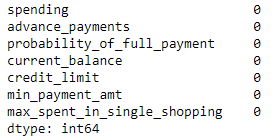
The dataset consist of 7 different attributes of credit card data. There are 210 entries, all datatypes are float type and no null value present in the dataset.

### **Summary of the Dataset**



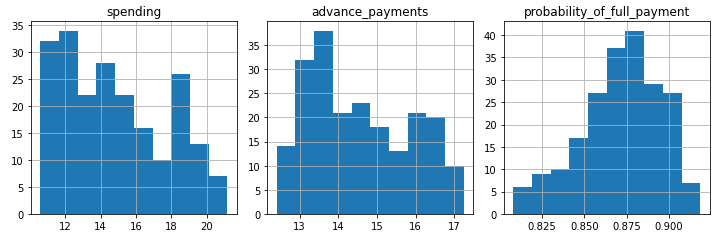
**Checking for Missing Values**

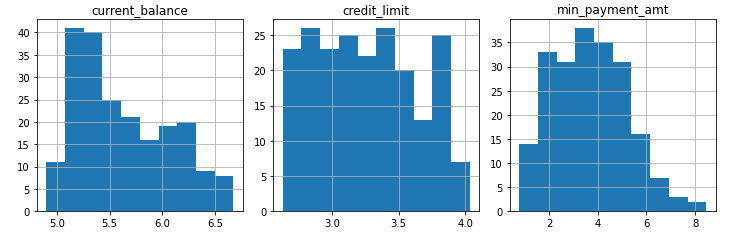
As computed from below, the dataset does not have any null or NA values.

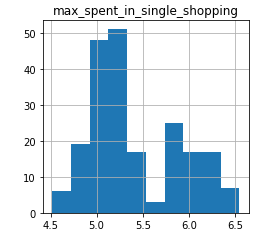


### **Univariate Analysis and Bivariate Analysis**

Histograms are plotted for all the numerical variables



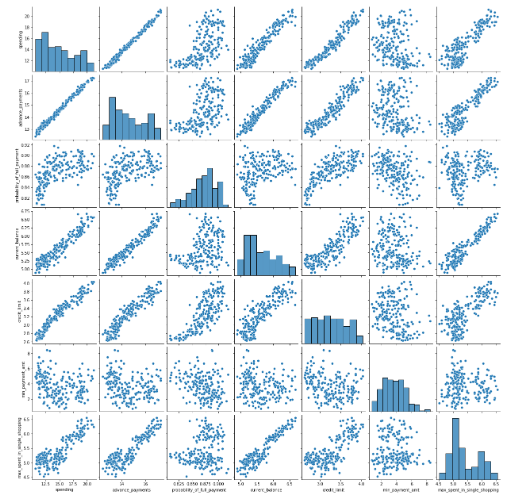




**Inference**: After plotting the Boxplots for all the variables we can conclude that a few outliers are present in the variable namely, **min\_payment\_amt** which means that there are only a few customers whose minimum payment amount falls on the higher side on an average. Since only one of the seven variable have a very small outlier value, hence there is no need to treat the outliers. This small value will not create any difference in our analysis.

We can conclude from the above graphs that the most of the customers in our data have a higher spending capacity, high current balance in their accounts and these customers spent a higher amount during a single shopping event. Majority of the customers have a higher probability to make full payment to the bank.

### **Pair Plot for all the variables**



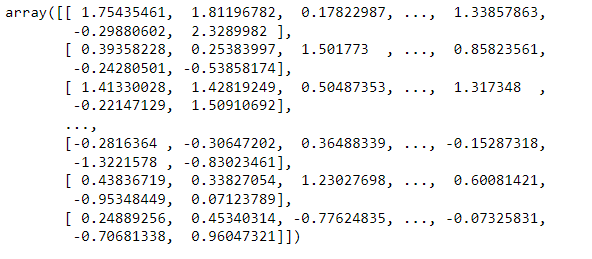
With the help of the above pair plot we can understand the Univariate and Bivariate trends for all the variables in the dataset.

## To provide justification whether scaling is necessary for clustering in this case

Feature scaling or Standardization is a technique for Machine Learning algorithms which helps in pre- processing the data. It is applied to independent variables which helps to normalise the data in a particular range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

For the data given to us, scaling is required as all the variables are expressed in different units such as spending in 1000’s, advance payments in 100’s and credit limit in 10000’s, whereas probability is expressed as fraction or decimal values. Since the other values expressed in higher units will outweigh probabilities and can give varied results hence it is important to Scale the data using Standard Scaler and therefore normalise the values where the means will be 0 and standard deviation 1.

Scaling of data is done using importing a package called StandardScaler from sklearn.preprocessing. For further clustering of dataset, we will be using the scaled data, “scaled\_bank\_df”.

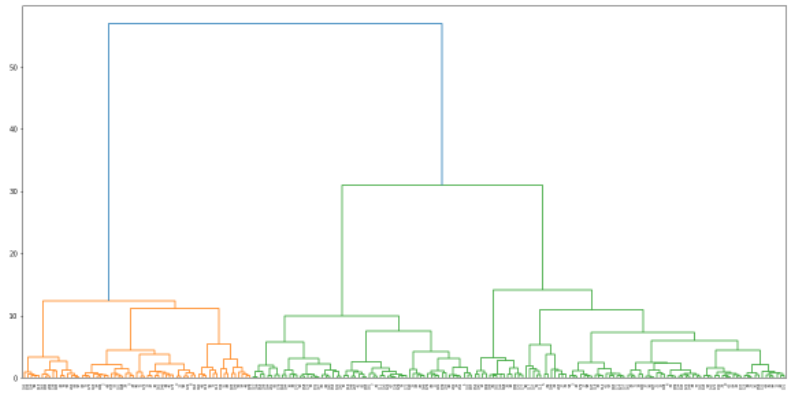


## To perform hierarchical clustering to scaled data and identify the number of optimum clusters using Dendrogram and briefly describe them.

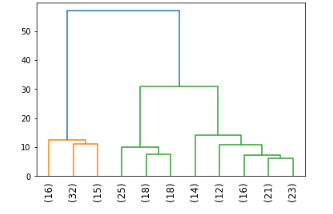
Cluster Analysis or Clustering is a widely accepted Unsupervised Learning technique in Machine Learning, Clustering can be divided into two categories namely, Hierarchical and K-means clustering.

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 similar to each other. There are two types of hierarchical clustering, Divisive and Agglomerative.

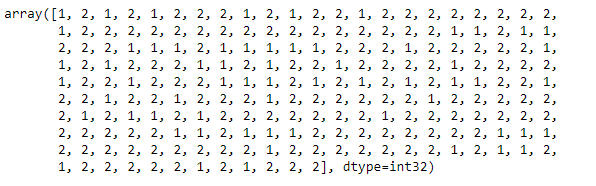
For the dataset in question we will be using Agglomerative Hierarchical Clustering method to create optimum clusters and categorising the dataset on the basis of these clusters.



We have created a dendrogram which shows two clusters very clearly. Now, we will check the make-up of these two clusters using ‘maxclust’ and ‘distance’. As can be seen from above we will now take 2 clusters for our further analysis.



This above graph shows the last 11 links in the dendrogram.



There are 2 optimum no of clusters.Cluster 1 consist of higher values of "Spending", "Max\_spend\_in\_single\_shopping", "advance\_payments", "credit\_limit", "current balance".

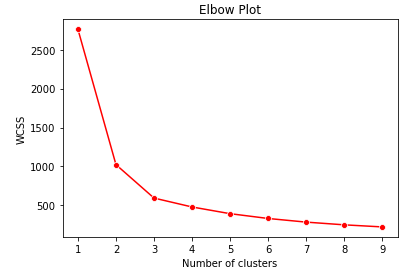
Clusters 2 consist of lower values of these attributes.

## To perform K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

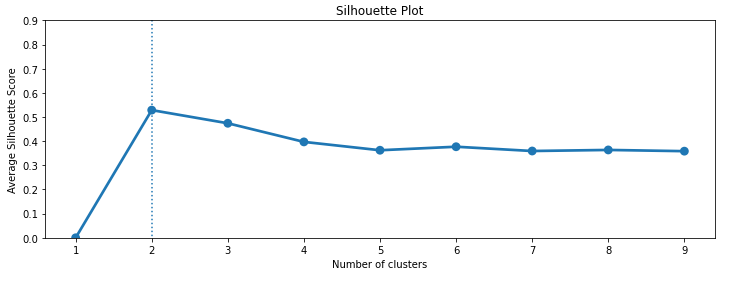
K-means clustering is one of the unsupervised machine learning algorithms. K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.

For the dataset we will be using K-means clustering on scaled data and identify the clusters formed and use them further to devise tools to target each group separately.

Firstly, we have scaled the dataset using StandardScaler package from sklearn.preprocessing and using the scaled\_bank\_df we will now plot two curves to determine the optimal number of clusters(k) to be used for our clustering. The two methods are namely within sum of squares(wss) method and average silhouette scores method.

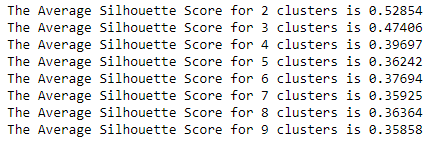


As per the above plot i.e. within sum of squares (wss) method we can conclude that the optimal number of clusters to be taken for k-means clustering is 3 since as per the elbow method it can be easily seen in the curve that after 3 the curve gets flat



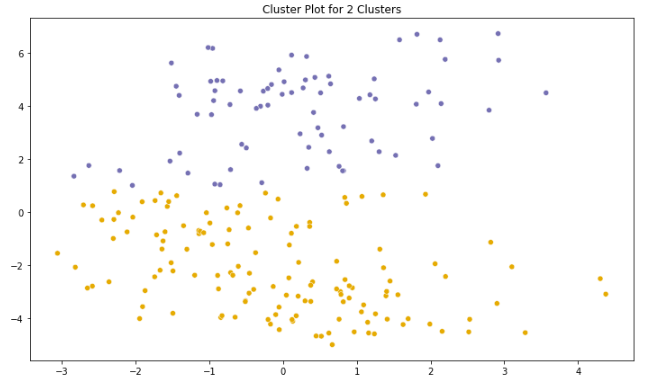
It is clear from above figure that the maximum value of average silhouette score is achieved for k =2, which, therefore, is considered to be the optimum number of clusters for this data. But statistically 2 clusters are not good for the analysis and does'nt full fill the need for clustering. Hence selecting 2 close optimum value of k other than 2.

The silhouette scores and silhouette widths are calculated using silhouette\_samples and silhouette\_score package from sklearn.metrics. The average silhouettes score is coming to be 0.400 and minimum silhouette score is 0.002. The silhouette score ranges from -1 to +1 and higher the silhouette score better the clustering.



## 1.5. To describe cluster profiles for the clusters defined and recommend different promotional strategies for different clusters.

Now, the final step is to identify the clusters that we have created using Hierarchical clustering and K- means clustering for our market segment analysis and devise promotional strategies for the different clusters. Since from the above analysis we have identified 2 clusters from hierarchical clustering and 3 optimal clusters from k-means clustering. We will now further analyse and determine the best clustering approach that can be helpful for the market segmentation problem in hand. We will first plot and map the clusters from both the methods.



Now, in the below table we have tabulated the averages for all the variables of the five clusters created from the above clustering using hierarchical and K-means methods. As per the values we can segment the clusters into two for Hierarchical and three segments for K-means clusters.

#### Segments

**Hierarchical Cluster 1**: This segment has higher spending per month, high current balance and credit limit. This is the **Prosperous or Upper class** with majorly higher income. This segment can be targeted using various offers such as cards with rewards and loyalty points for every spent.

**Hierarchical Cluster 2**: This segment has must lower spending per month with low current balance and lower credit limit. This is the **Middle Class** with low incomes. This segment can be targeted with cards that have lower interest rates so as to encourage more spending.

**K-means Cluster 0**: This segment has the lowest spending per month, lowest current balance and credit limit. This is the **Financially Stressed Class** with very low income on an average. This segment can be targeted with cards with offers such as zero annual charges and lurking them with benefits such as free coupons or tickets and waivers on a variety of places.

**K-means Cluster 1**: This segment has higher spending per month, high current balance and credit limit. This is the **Prosperous or Upper class** with majorly higher income. This segment can be targeted using various offers such as cards with rewards and loyalty points for every spent.

**K-means Cluster 2**: This segment has must lower spending per month with low current balance and lower credit limit. This is the **Middle Class** with low incomes. This segment can be targeted with cards that have lower interest rates so as to encourage more spending.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **VARIABLES** | **Spending** | **Advance Payments** | **Probabilit y of full payment** | **Current balance** | **Credit Limit** | **Min Payment Amt** | **Max spent in single**  **shopping** |
| **Hierarchical**  **(Cluster 1)** | 18.62 | 16.26 | 0.88 | 6.19 | 3.71 | 3.66 | 6.06 |
| **Hierarchical**  **(Cluster 2)** | 13.23 | 13.83 | 0.87 | 5.39 | 3.07 | 3.71 | 5.13 |
| **K-means**  **(Cluster 0)** | 11.86 | 13.25 | 0.85 | 5.23 | 2.85 | 4.74 | 5.1 |
| **K-means**  **(Cluster 1)** | 18.5 | 16.2 | 0.88 | 6.18 | 3.7 | 3.63 | 6.04 |
| **K-means**  **(Cluster 2)** | 14.43 | 14.33 | 0.88 | 5.51 | 3.26 | 2.7 | 5.12 |

PROBLEM 2: CART-RF-ANN

The dataset provided to us is stored as “insurance\_part2\_data.csv” which contains data of 3000 customers and 10 variables namely:

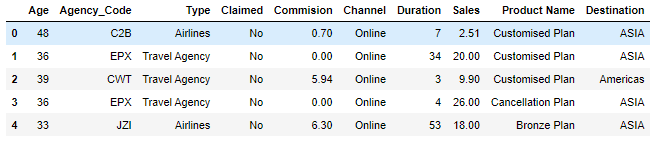
|  |  |
| --- | --- |
| Age | Age of insured |
| Agency\_Code | Code of tour firm |
| Type | Type of tour insurance firms |
| Claimed | Target: Claim Status |
| Commission | The commission received for tour insurance firm |
| Channel | Distribution channel of tour insurance agencies |
| Duration | Duration of the tour |
| Sales | Amount of sales of tour insurance policies |
| Product Name | Name of the tour insurance products |
| Destination | Destination of the tour |

# SOLUTIONS

## 2.1 To read the dataset and perform the descriptive statistics and do null value condition check and write an inference on it.

**Importing the Dataset**

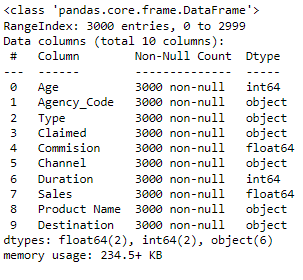
The dataset in question is imported in jupyter notebook and will store the dataset in “**insurance\_df**”. The top 5 rows of the dataset are viewed.

****

**Dimension of the Dataset**

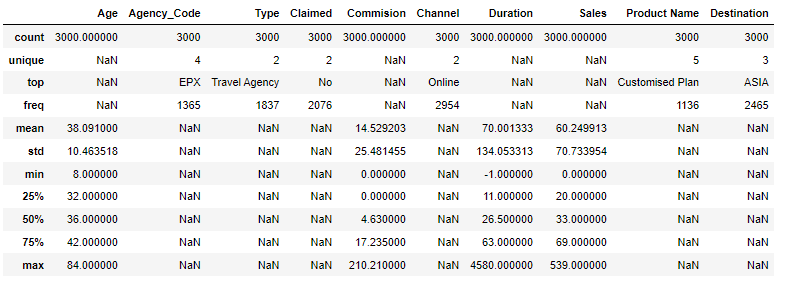
****

**Structure of the Dataset**

****

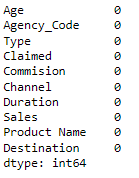
* Claimed is the target variable while all others are the predictors
* Out of 9 datatypes 2 are integer type, 2 are float and 6 are object types
* It seems there is no null values present in dataset

### **Summary of the Dataset**



### **Checking for Missing Values**

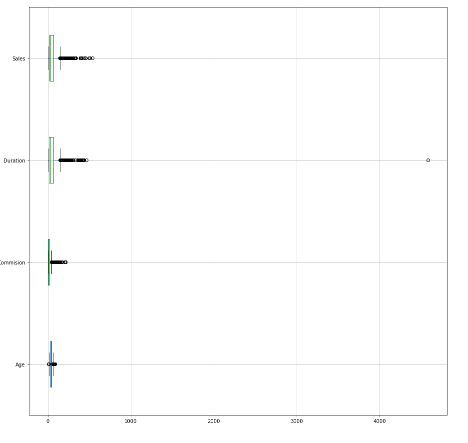
The missing values or “NA” needs to be checked and dropped from the dataset for the ease of evaluation and null values can give errors or disparities in results.



**Dropping the non-important columns**

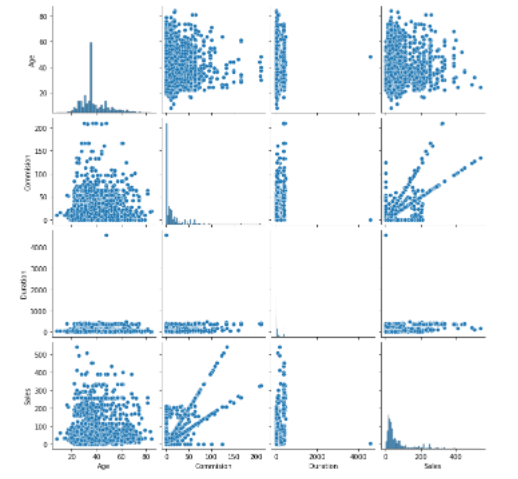
In this dataset, “Agency\_Code” Is the column which cannot be used for our analysis. Hence, we will be dropping this column

**Checking for outliers.**

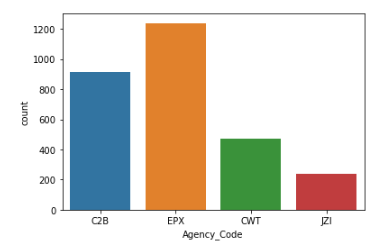
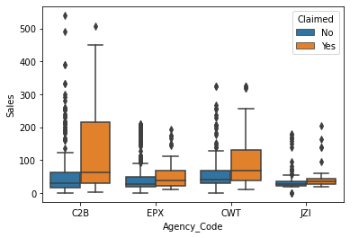
****

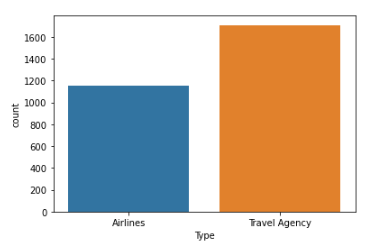
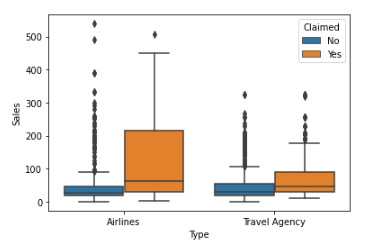
**Inference:** After plotting the Boxplots for all the variables we can conclude that a few outliers are present in the variable namely, min\_payment\_amt which means that there are only a few customers whose minimum payment amount falls on the higher side on an average. Since only one of the seven variable have a very small outlier value, hence there is no need to treat the outliers. This small value will not create any difference in our analysis. We can conclude from the above graphs that the most of the customers in our data have a higher spending capacity, high current balance in their accounts and these customers spent a higher amount during a single shopping event. Majority of the customers have a higher probability to make full payment to the bank.

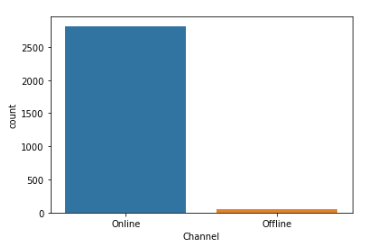
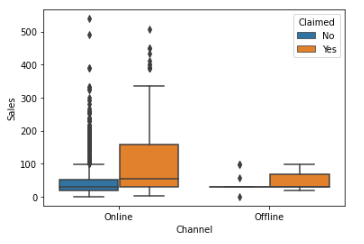
**Checking pairwise distribution of the continuous variables:**

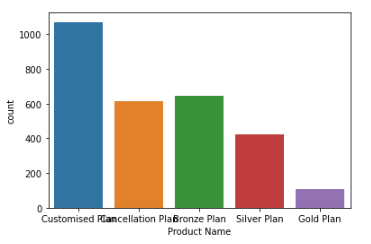
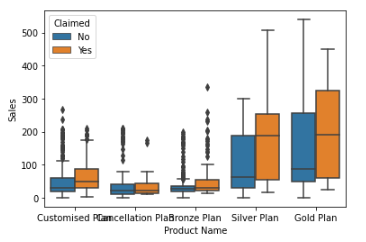
****

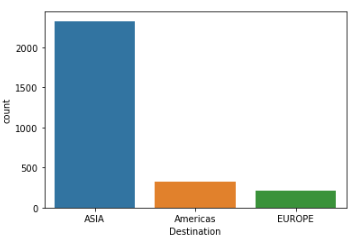
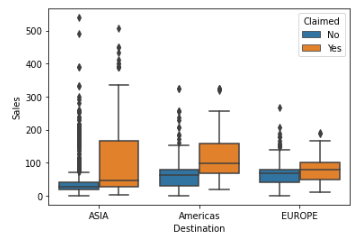
**Categorical Variables:**

**** ****

**** ****

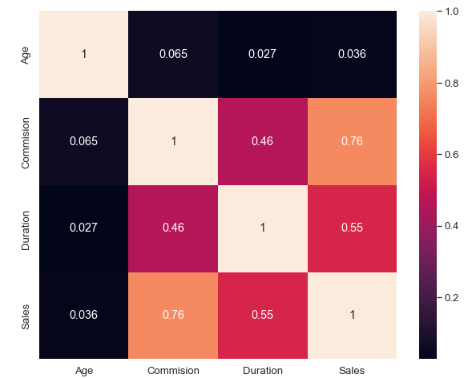
**** ****

**** ****

**** ****

**Checking for Correlations:**

This graph can help us to check for any correlations between different variables.

****

As interpreted from the above heat map, there is no or extremely low correlation between the variables given in the dataset. There are mostly positive correlation between different attributes. Only the "Sales" & "Commission" are highly correlated.

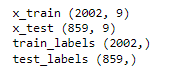
## 2.2. To split the data into test and train, build classification model CART, Random Forest and Artificial Neural Network.

## 

For our analysis and building Decision tree and Random Forest, we have to convert the variables which have ‘object’ datatype and convert them into integer.

**Splitting Dataset in Train and Test Data (70:30)**

For building the models we will now have to split the dataset into Training and Testing Data with the ratio of 70:30. These two datasets are stored in X\_train and X\_test with their corresponding dimensions as follows:

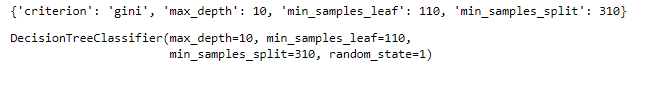
****

## CART Model

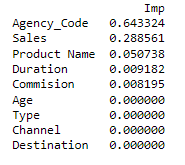
Classification and Regression Trees(CART) are a type of Decision trees used in Data mining. It is a type of Supervised Learning Technique where the predicted outcome is either a discrete or class (classification) of the dataset or the outcome is of continuous or numerical in nature(regression). Using the Train Dataset(X\_train) we will be creating a CART model and then further testing the model on Test Dataset(X\_test)

With the help of DecisonTreeClassifier we will create a decision tree model and using the “gini” criteria we will fit the train data into this model. After this using the tree package we will create a dot file namely, claim\_tree.dot to help visualize the tree. Below are the variable importance values or the feature importance to build the tree

Using the GridSearchCV package from sklearn.model\_selection we will identify the best parameters to build a regularised decision tree. Hence, doing a few iterations with the values we got the best parameters to build the decision tree which are as follows



These best grid parameters are henceforth used to build the regularised or pruned Decision tree.

****

Looking at the above important parameters the model higly depends upon at "Agency Code" i.e **64.33%** and "Sales" i.e **28.89%**

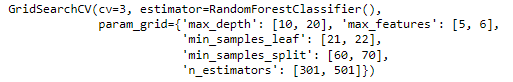
The regularised Decision tree was formulated using best grid parameters computed above and with the “gini” criteria it is fitted in the train dataset. The regularised tree is stored as a dot file namely, claim\_tree\_regularised.dot and can be viewed using webgraphviz in the browser.

## Random Forest

Random Forest is another Supervised Learning Technique used in Machine Learning which consists of many decision trees that helps in predictions using individual trees and selects the best output from them.

Using the Train Dataset(X\_train) we will be creating a Random Forest model and then further testing the model on Test Dataset(X\_test) For creating the Random Forest, the package “RandomForestClassifier” is imported from sklearn.metrics.

Using the GridSearchCV package from sklearn.model\_selection we will identify the best parameters to build a Random Forest namely, rfcl. Hence, doing a few iterations with the values we got the best parameters to build the RF Model which are as follows





Using these best parameters evaluated using GridSeachCV a Random Forest Model is created which is further used for model performance evaluation.

## Artificial Neural Network (ANN)

Artificial Neural Network(ANN) is a computational model that consists of several processing elements that receive inputs and deliver outputs based on their predefined activation functions. Using the train dataset(X\_train) and test dataset(X\_test) we will be creating a Neural Network using MLPClassifier from sklearn.metrics. Firstly, we will have to Scale the two datasets using Standard Scaler package.

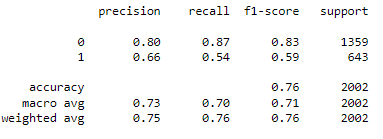
## To check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model.

To check the Model Performances of the three models created above certain model evaluators are used i.e., Classification Report, Confusion Matrix, ROC\_AUC Score and ROC Plot. They are calculated first for train data and then for test data.

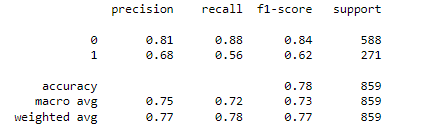
**CART Model**

**Classification Report**

**Train:**

****

**Test:**

****

**Confusion Matrix**

**Train:**

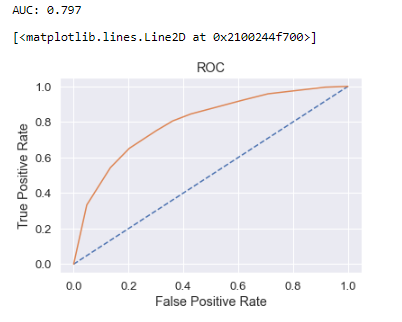
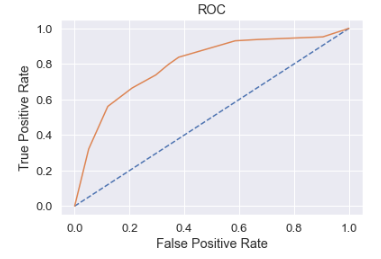
****

**Test:**

****

### **OC\_AUC Score and ROC Curve**

Train: Test:

**Model Score**

Train:

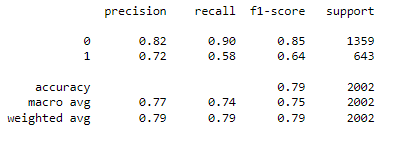
0.7622377622377622

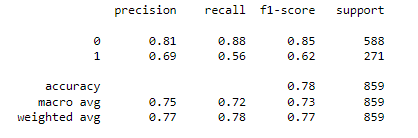
Test:

0.7788125727590222

**Random Forest Model**

**Classification Report**

****Train: Test:



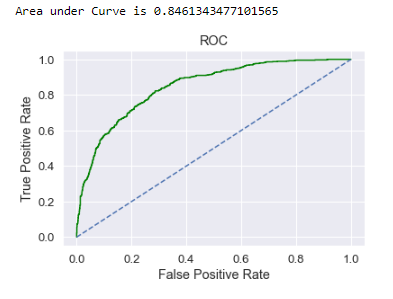
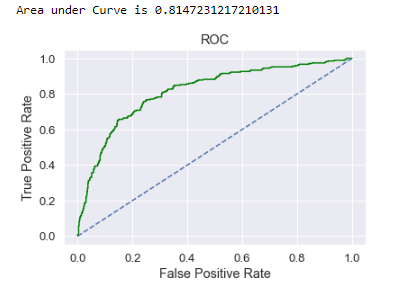
**Confusion Matrix**

Train: Test:

**AUC\_ROC Score and ROC Curve**

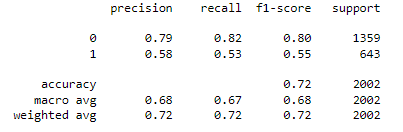
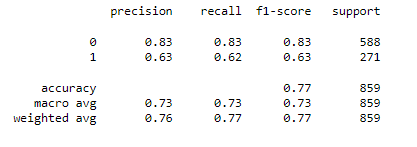
Train: Test:

**** ****

**Artificial Neural Network Model**

**Classification Report**

**Train: Test:**

** **

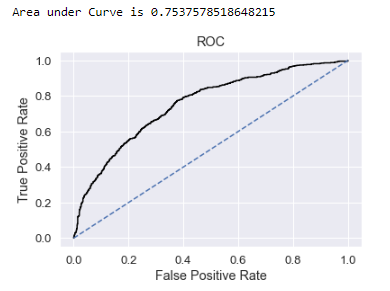
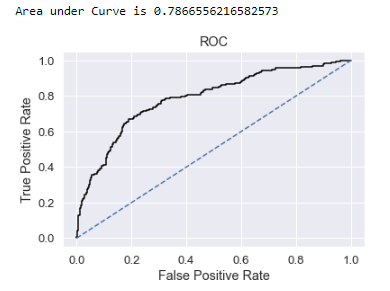
**Confusion Matrix**

**Train: Test:**

** **

**AUC\_ROC Score and ROC Curve**

**Train: Test:**

** **

## To compare all the models and write an inference which model is best/optimized.

Comparison of all the performance evaluators for the three models are given in the following table. We are using Precision, F1 Score and AUC Score for our evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **F1 Score** | **AUC Score** |
| **CART Model** |  |  |  |
| *Train Data* | 0.66 | 0.60 | 0.81 |
| *Test Data* | 0.68 | 0.60 | 0.80 |
| **Random Forest** |  |  |  |
| *Train Data* | 0.72 | 0.64 | 0.84 |
| *Test Data* | 0.68 | 0.58 | 0.82 |
| **Neural Network** |  |  |  |
| *Train Data* | 0.58 | 0.55 | 0.75 |
| *Test Data* | 0.63 | 0.63 | 0.79 |

### **Insights:**

From the above table, comparing the model performance evaluators for the three models it is quite clear that the Random Forest Model is performing well as compared to the other two as it has high precisions for both training and testing data and although the AUC Score is the same for all the three models for training data but for testing data it is the highest for Random Forest Model. Choosing Random Forest Model is the best option in this case as it will exhibit very less variance as compared to a single decision tree or a multi – layered Neural Network.

## To provide business insights and recommendations.

For the business problem of an Insurance firm providing Tour Insurance, we have attempted to make a few Data Models for predictions of probabilities. The models that are attempted are namely, CART or Classification and Regression Trees, Random Forest and Artificial Neural Network(MLP). The three models are then evaluated on training and testing datasets and their model performance scores are calculated.

The Accuracy, Precision, F1 Score are computed using Classification Report. The confusion matrix, AUC\_ROC Scores and ROC plot are computed for each model separately and compared. All the three models have performed well but to increase our accuracy in determining the claims made by the customers we can choose the Random Forest Model. Instead of creating a single Decision Tree it can create a multiple decision trees and hence can provide the best claim status from the data.

As seen from the above model performance measures, for all the models i.e. CART, Random Forest and ANN have performed exceptionally well. Hence, we can choose either of the models but choosing Random Forest Model is a great option as even though they exhibit the same accuracy but choosing Random Forest over Cart model is way better as they have much less variance than a single decision tree.

\*-\*-\*-\*-\*-\*-\*