DATA MINING PROJECT

ANSWER REPORT

SULOCHANA

**Problem 1: Clustering**

A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. **Please note that it is a summarized data that contains the average values in all the columns considering all the months, and not for any particular month**. You are given the task to identify the segments based on credit card usage.

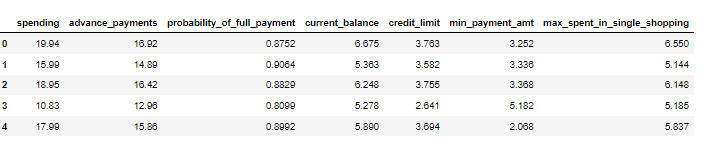
**Data** **Dictionary** **for** **Market** **Segmentation:**

1. spending: Amount spent by the customer using the credit card per month (in 1000s). For example, if the spending is 19.94, then the customer has actually spent (19.94 \* 1000 = 19940) 19940 Rs per month on an average.
2. advance\_payments: Amount paid by the customer in advance by cash even before the credit card bill got generated for any particular month (in 100s). For example, if the advance\_payments is 16.92, then the customer has paid (16.92\*100 = 1692) 1692 Rs on an average per month.
3. probability\_of\_full\_payment: Probability of the credit card payment done in full by the customer to the bank. If it is 0.8752, then it means that the customer has a chance of 87.52% to pay the entire credit card bill on an average per month.
4. current\_balance: The balance amount left in the credit card account to make the future purchases (in 1000s). For example, if the current\_balance is 6.675, then it means that the customer is left out with a credit card balance of (6.675\*1000 = 6675) 6675 Rs which he can use for the future purchases.
5. credit\_limit: Limit of the amount in credit card (10000s) sanctioned by the bank to the customer. For example, if the credit\_limit is 3.763, it means that the customer has been sanctioned a credit card limit of (3.763\*10000 = 37,630) 37630 Rs.
6. min\_payment\_amt : The average minimum amount paid by the customer while making payments for the credit card bill purchases made monthly (in 100s). For example, if the min\_payment\_amt is 3.252, it means that the customer has paid only (3.252\*100 = 325.2) 325.2 Rs as the minimum payment instead of paying the entire credit card bill amount on an average per month.
7. max\_spent\_in\_single\_shopping: Maximum amount spent by the customer for a single transaction using the credit card (in 1000s). For example, if the max\_spent\_in\_single\_shopping is 6.55, it means that the customer has spent a maximum of (6.55\*1000=6550) 6550 Rs for a single transaction using credit card on an average per month.

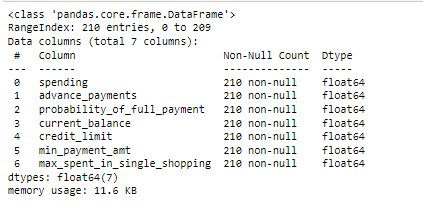
The dataset imported in jupyter notebook by using pd\_read.csv function and stored in object called ‘df’**.** Then, check wether the data set loaded properly or not by using df.head() function. Here is the sample data. It indicates data loaded perfectly.

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

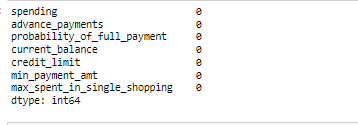
**SAMPLE OF THE DATA**



**Data info:**

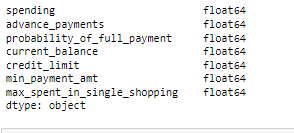


Based on above output this data set has 7 variables and 210 observations.



Based on above output there is no null values in data set.

**Data types:**

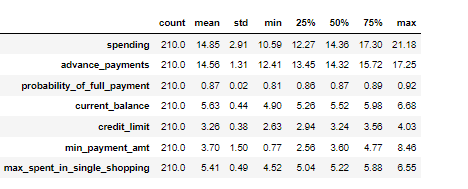


Above table shows all the variables in the dataset are numeric variables.

**Duplicates:**

There are no duplicates in the dataset.

**Data describe:**



The above table shows 5 point summary of the data.

Based on output the data looks good.

All the variables mean and median are nearly equal.

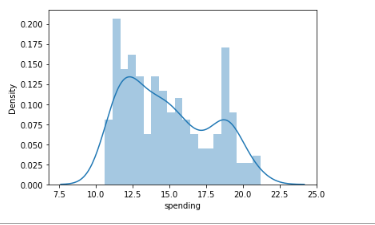
Here we can see all the variables minimum and maximum are as would expect.

But the minimum and maximum of min\_payment\_amt seem to be a large difference.

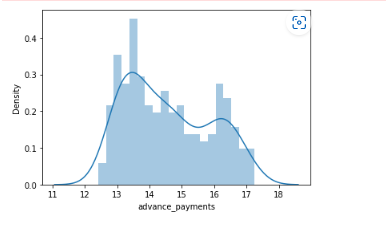
Before going to the further analysis, we confirmed if we have any duplicate values in the data. There are no duplicate values in the dataset.

**Univariate Analysis:**

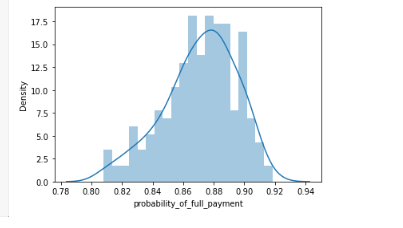
**Spending:**



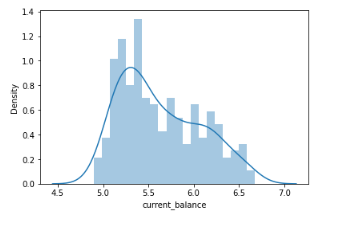
**Advance\_payments:**



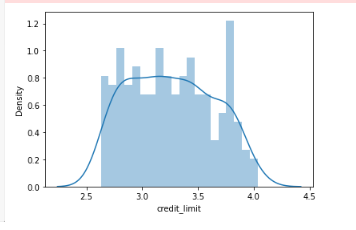
**Probability \_of\_full\_payments:**



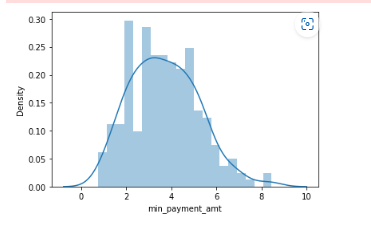
**Current\_balance:**



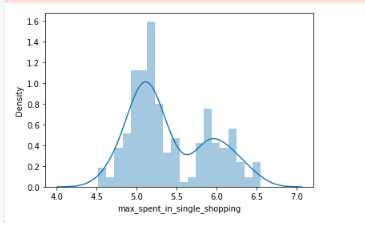
**Credit\_limit:**



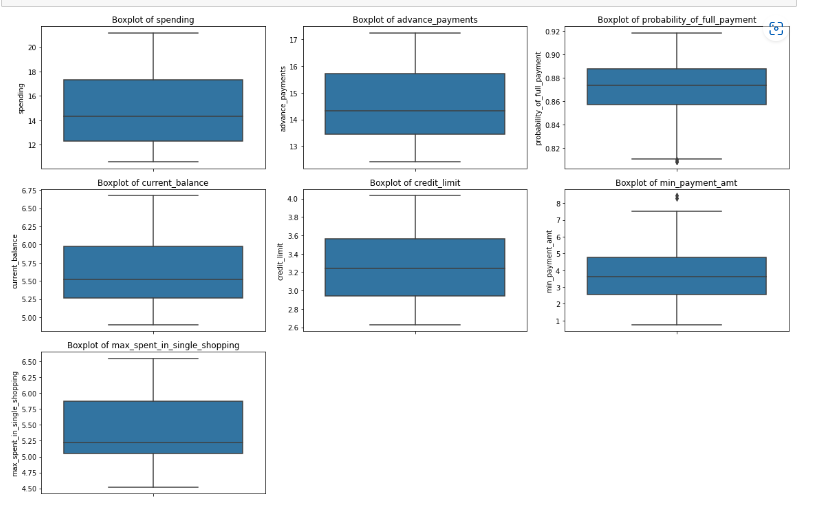
**Min\_payment\_amt:**



**Max\_spent\_in\_single\_shopping:**



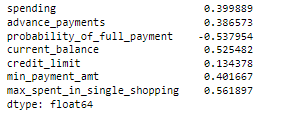
Box plots are used to identify the outliers in the data.



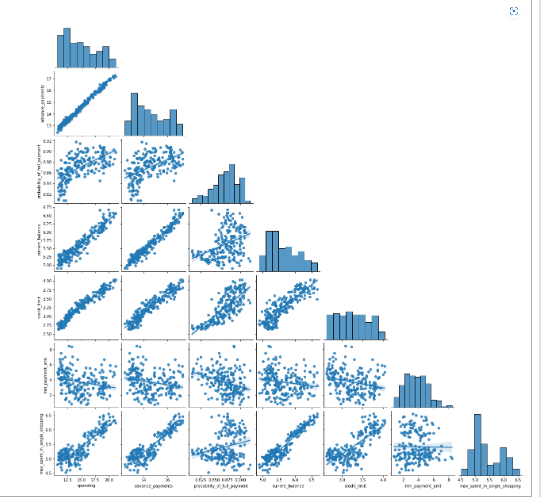
**Based on above output:**

* We can say that the spending parameter is right skewed.
* The advance\_payments parameter seems to be right skewed.
* we can observe say that the probability of full payment parameter is left skewed
* The current balance parameter seems to be right skewed.
* The credit\_limit parameter seems to be normal distribution.
* We can say that the min\_payment\_amt parameter is normally distributed.
* From above figure, we can say that the max\_spent\_in\_single\_shopping parameter is right skewed.
* There are outliers in both probability\_of\_full\_payment and min\_payment\_amt.

Let’s have a look at variable skewness in the next code.



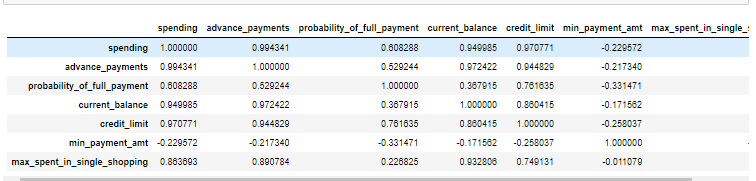
**Bivariate analysis/ Multivariate analysis:**

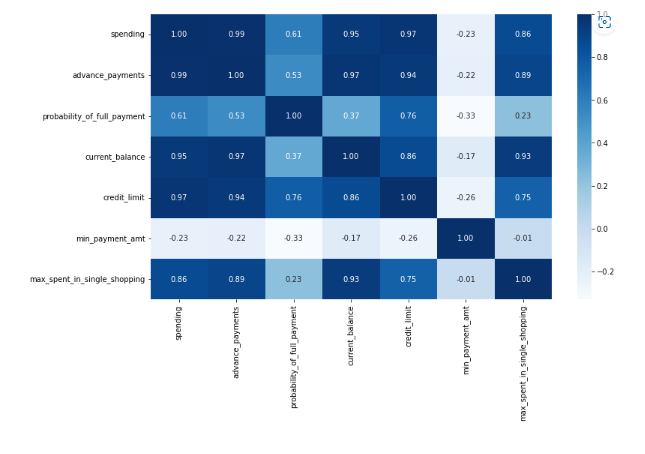


From above ouput, we can observe:

* Strong positive correlation between
* spending & advance\_payments
* advance\_payments & current\_balance
* credit\_limit & spending
* spending & current\_balance
* credit\_limit & advance\_payments
* max\_spent\_in\_single\_shopping and current\_balance

we have seen high correlations in between variables here, let’s have a look at the correlation table and the heatmap for more clarity.



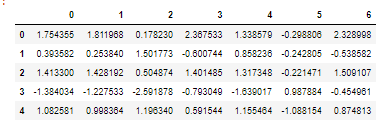


**Do you think scaling is necessary for clustering in this case? Justify**

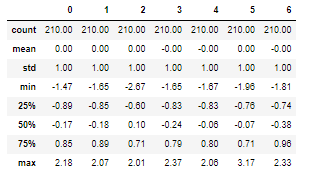
Yes.scaling is necessary for clustering in this case as standardising the data prevents the variability of the dataset, it convert data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

Here, I used standardscaler function from skeleran.preprocessing library.

The head of the scaled data:

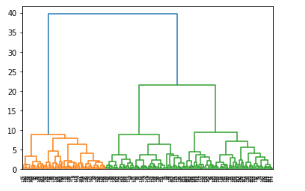


Lets confirme the data is scaled properly or not by using describe fuction. Here,we can see mean and std for all variables are 0 and 1. Hence, we can confirm the data is scaled.

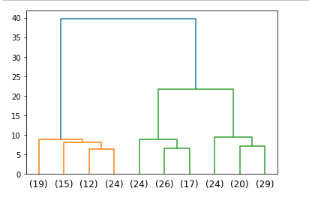


. Here,we can see mean and std for all variables are 0 and 1. Hence, we can confirm the data is scaled.

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

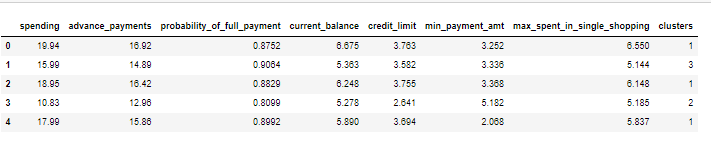


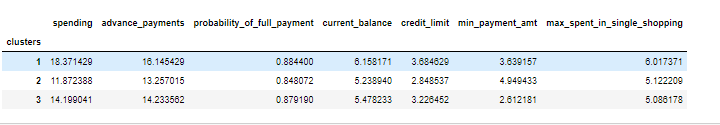
Here I applied hierarchical clustering to the scaled data. I have used dendrogram, linakge function from scipy.cluster.hierarchy library. But here we can not observe clusters division clearly. So lets have a look closly by using truncate\_ mode last 10 clusters.

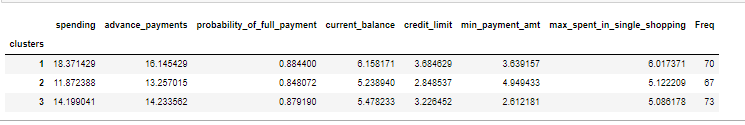


Hierarchical clustering is also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly like each other.

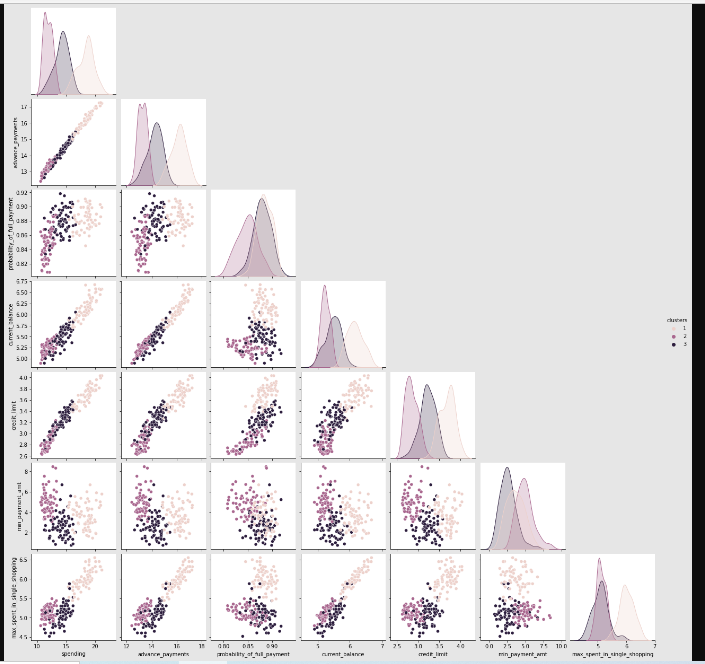
After performing hierarchical clustering on scaled data, we obtain mean values within 3 cluster formations as follows:





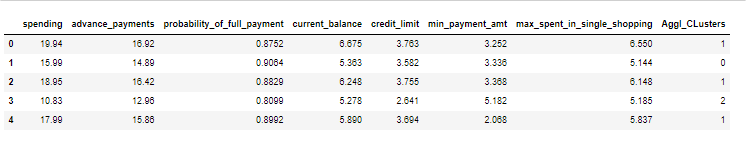


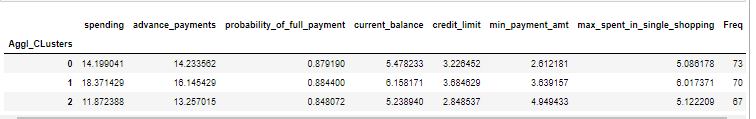
We can see a visual representation of the clustering applied on the dataset by implot.



Now we will perform Agglomerative Clustering on the scaled data:

Below is the sample data after performing Agglomerative clustering and also Frequency distribution as per the agglomerative clustering on the mean values of the dataset:





We can say after performed the hierarchical clustering based on dendrogram the optimum number of clusters are 3.

As we can observe from the 3 cluster segmentations, the customers under the high spenders cluster have higher valuations and probabilities across the various criteria mentioned except the min\_payment\_amt where the customers of the low spenders cluster have a higher bill value amount as their minimum amount that would have to be remitted.

## Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called **flat clustering** algorithm. The number of clusters identified from data by algorithm is represented by ‘K’ in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

[1470.0000000000002,

659.1717544870406,

430.6589731513006,

371.2935481943965,

327.4472622369584,

289.33066945965584,

264.1397930395344,

243.8660295853418,

223.1040691610046,

206.78834352214284]

## 

## After performing k - means clustering on scaled data, according to scree plot 3 clusters are enough to consider.

## 

## Silhouette score:

Silhouette score is used to evaluate the quality of clusters created using algorithms K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters. To calculate the Silhouette score for each observation/data point, the following distances need to be found out for each observations belonging to all the clusters:

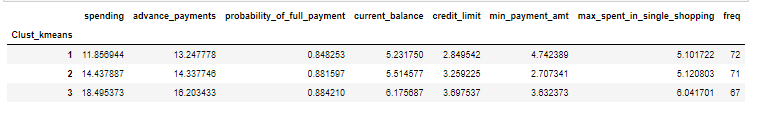
The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

Silhouette score:

0.4007270552751298

However, the Silhouette Score of clusters was more appropriate; however, objective 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.

The mean values for 3 clusters:



Here, we can recommend optimum number of clusters as 3 for Business Recommendations.

Cluster 1: high spending group

* For this we can offer higher reward points on a higher probability can increase their spending capacity.
* Providing low cost EMI as a promotional scheme for some popular bands on online shopping can be a great motivator for this group.
* 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.

Cluster 2: 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 marginalized 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.

Cluster 3: Low Spending Group

• We can spend some time analyzing 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 these 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.

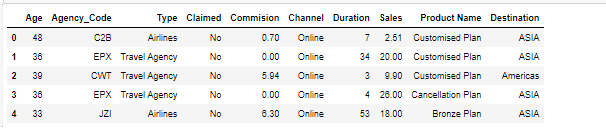
**Attribute** **Information:**

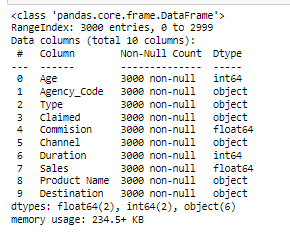
1. Target: Claim Status (Claimed)  
   2. Code of tour firm (Agency Code)  
   3. Type of tour insurance firms (Type)  
   4. Distribution channel of tour insurance agencies (Channel)  
   5. Name of the tour insurance products (Product)  
   6. Duration of the tour (Duration in days)  
   7. Destination 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)

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

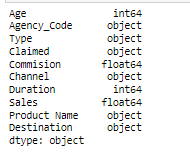
The dataset imported in jupyter notebook by using pd\_read.csv function and stored in object called ‘df2’**.** Then, check whether the data set loaded properly or not by using df2.head () function. Here is the sample data. It indicates data loaded perfectly.

SAMPLE OF THE DATA



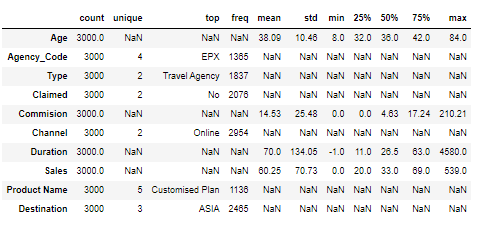


* In this dataset total 3000 observations and 10 variables are there.
* There are no null values in this dataset.



Above output shows, in this dataset the variables age, commission, duration, sales are numeric data type and agency \_code, type, claimed, channel, product name, destination are object type variables.

**Descriptive statistics:**

****

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

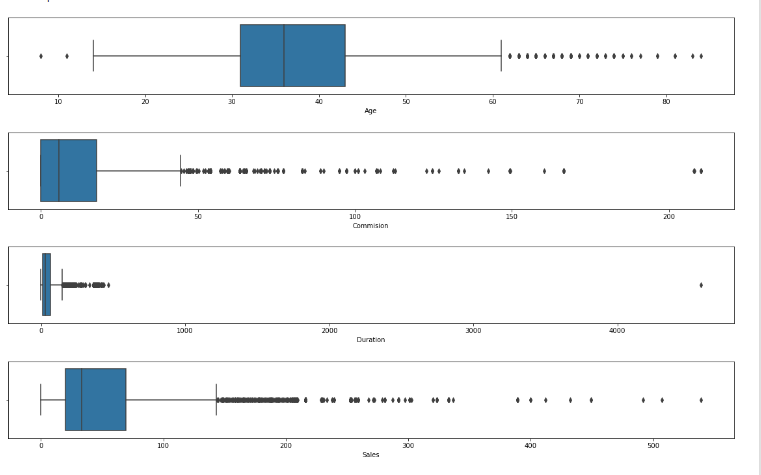
▪ Product Name is Customized Plan with a frequency of 1136

▪ Destination is ASIA with a frequency of 2465

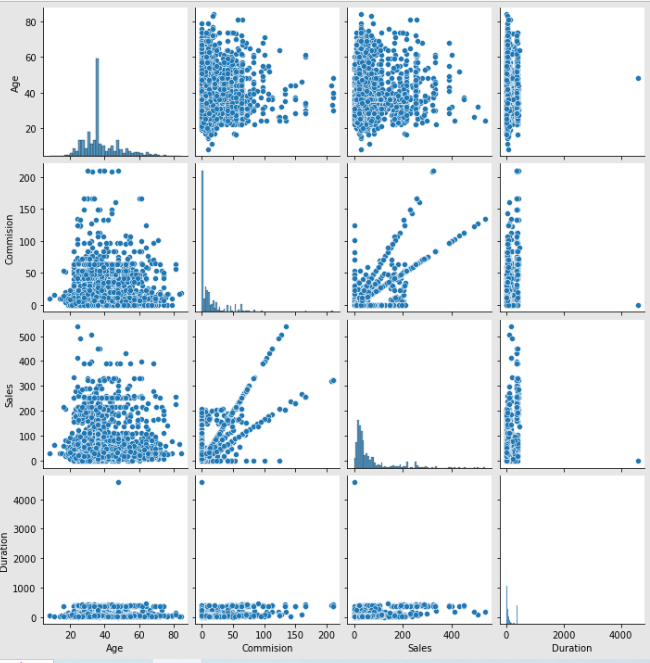
* For the float and integers data type values like: Age, Commission, Duration and Sales the difference between its 75th percentile and Max value is very large, indicating there will be large number of outliers in the data.

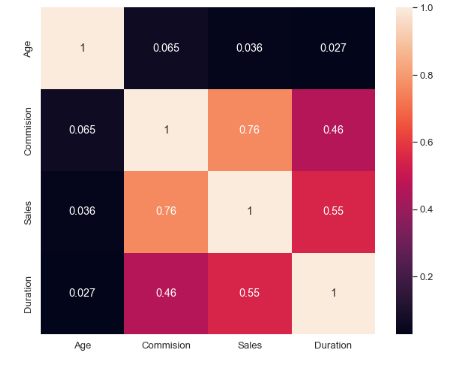
**Duplicates**:

There are total 139 duplicates in this data.



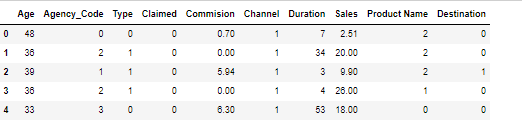
There are multiple outliers in the data. However, since the outliers do not directly impact any of the three models, treating the outliers is not necessary at this stage.





There is no major correlation in any of the two variables but in comparison, Sales and Commission has a correlation of 0.76 which is high in comparison with other variables.

Converting all object data type variables to Categorical Codes



The Proportions of 0s and 1st of our target variable:



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

Here splitting the data into train and test data by using train\_test\_split function from sklearn.model\_selection.

Below is the output:

X\_train (2002, 9)

X\_test (859, 9)

train\_labels (2002,)

test\_labels (859,)

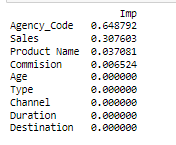
**CART**

A Classification and Regression Tree (CART), is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is a split in a predictor variable and each end node contains a prediction for the outcome variable.

Using the Decision Tree Classifier and the Grid search method, I have identified the best grid:

{'criterion': 'gini', 'max\_depth': 4, 'min\_samples\_leaf': 25, 'min\_samples\_split': 300}

From the decision tree, we extracted the variable importance shown below:



Based on above extract, Agency code is the most important variable in the dataset, followed by Sales and Product Name.

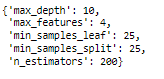
Commission has comparatively very less importance. However Age, Type, Channel, Duration and Destination have no importance in the model building.

**RANDOM FOREST**

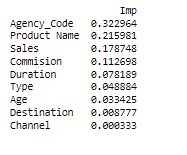
Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Using the random forest classifier and grid search function, we identified the best grid parameters:



From the random forest model, we extracted the variable importance as per RF:



Like CART, for RF as well Agency\_code has the most importance in the model, however Sales and Product Name exchanged places. In this model, each of the variables plays a role in model building at some importance level but Channel Variable has the lowest importance of them all.

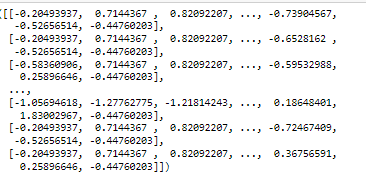
**NEURAL NETWORK CLASSIFIER**

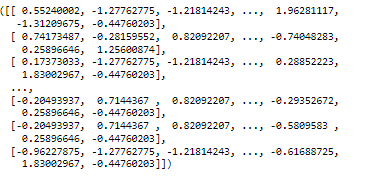
Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record.

Neurons are organized into layers: input, hidden and output. The input layer is composed not of full neurons, but rather consists simply of the record's values that are inputs to the next layer of neurons. The next layer is the hidden layer. Several hidden layers can exist in one neural network. The final layer is the output layer, where there is one node for each class.

For this model we have to scale the data. Here using standardscaler function for scaling the data.

The below output shows sample data after scaling.





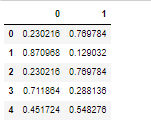
Using the MLP Classifier and grid search, I identified the best grid:

MLP Classifier(hidden\_layer\_sizes=8, max\_iter=2500, random\_state=0, tol=0.001)

**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.**

**CART Performance Matrix:**

After predicting the test and train data, below is the head of ytest\_predict\_prob:

****

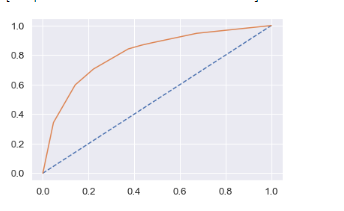
**Train Data Accuracy:**

0.7772227772227772

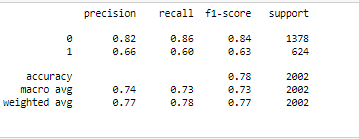
**Train data AUC score:**

AUC: 0.809

**AUC and ROC curve of CART model on Training Data**

****

**Confusion matrix for train data:**



cart\_train\_precision: 0.82

cart\_train\_recall: 0.86

cart\_train\_f1: 0.84

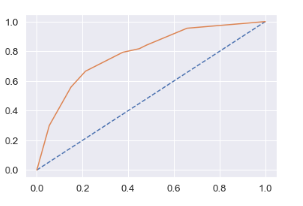
**AUC and ROC curve of CART model on Test Data**

**Test Data Accuracy:**

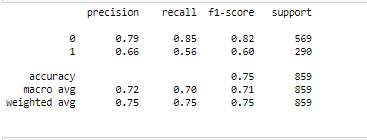
0.7520372526193247

**Test data AUC score:**

AUC: 0.786



**Confusion Matrix on CART Test data**



cart\_test\_precision: 0.79

cart\_test\_recall: 0.85

cart\_test\_f1: 0.82

**RANDOM FOREST ROC, AUC CURVE AND CONFUSION MATRIX**

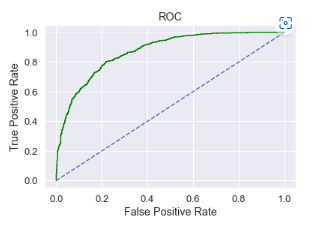
**RF Model evaluation on Train Data:**

**Train data accuracy:**

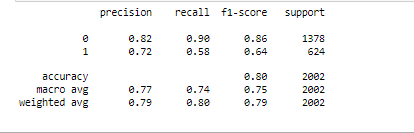
0.7972027972027972

**Train data AUC score:**

Area under Curve is 0.8499067303040453

****

**Train data confusion matrix:**



rf\_train\_precision: 0.82

rf\_train\_recall: 0.9

rf\_train\_f1: 0.86

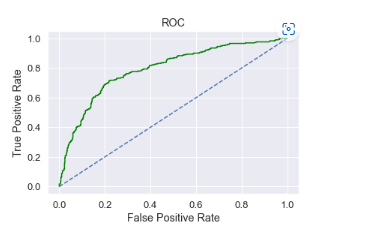
**RF Model evaluation on Test Data**

**Test data accuracy:**

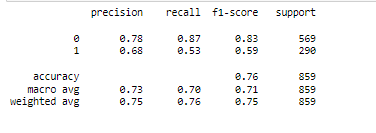
0.7566938300349243

**Test data AUC score:**

Area under Curve is 0.795354826980183

****

**Confusion matrix for test data:**



rf\_test\_precision: 0.78

rf\_test\_recall: 0.87

rf\_test\_f1: 0.83

**Neural Network Classifier:**

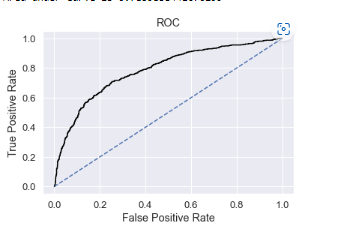
**NN Model evaluation on Training Data**

**Train data accuracy:**

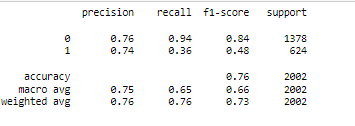
0.7597402597402597

**Train data AUC score:**

Area under Curve is 0.7859838441070299

****

**Confusion matrix for train data:**

****

nn\_train\_precision: 0.76

nn\_train\_recall: 0.94

nn\_train\_f1: 0.84

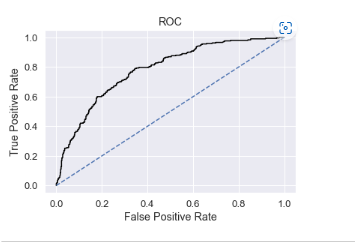
**NN Model evaluation on Test Data**

**Test data accuracy:**

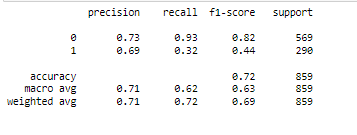
0.7217694994179278

**Test data AUC score:**

Area under Curve is 0.7810556935943276

****

**Confusion matrix for test data:**

****

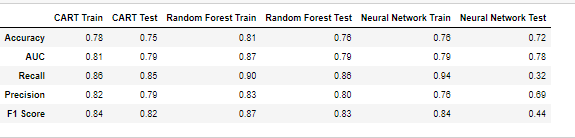
nn\_test\_precision: 0.69

nn\_test\_recall : 0.32

nn\_test\_f1: 0.44

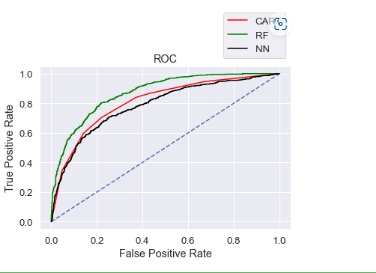
### Final Model: Compare the entire model and write an inference which model is best/optimized.

Now, we have to compare the three models accuracy, AUC, recall, precision and F1 score of all the models to know the best model for this business issue.

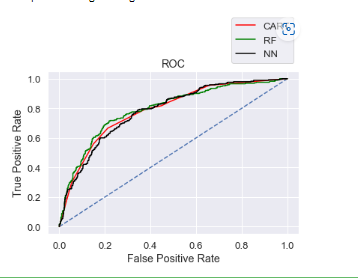


**ROC curve of the Train data of all the three models:**

ROC Curve for the 3 models on the Training data



ROC Curve for the 3 models on the Test data



* When we compared the three 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.

**After performing the whole analysis, below described the business insights and recommendations:**

* Streamlining online experiences benefitted customers, leading to an increase in conversions, which subsequently raised profits.
* In this data 90% of insurance is done by online channel.
* Almost all the offline business has a claimed associated.
* Need to train the JZI agency resources to pick up sales as they are in bottom, need to run promotional marketing campaign or evaluate if we need to tie up with alternate agency
* 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.
* 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.
* Since the variable Agency code seems 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.
* Team can easily target the customers who will not claim for tour insurance. Once Team receives customer data that 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.
* Product plan which has higher commission rate can be recommended to the set customers who will fall under NO claim status.

**END**