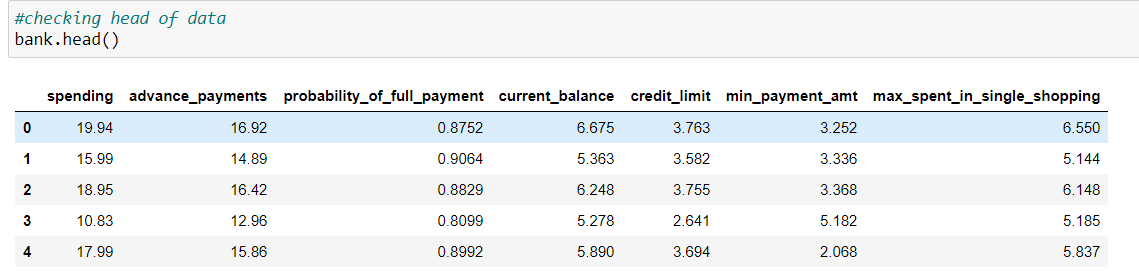
**Problem 1: Clustering**

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

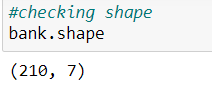
**Solution:-** lets check head of data:-



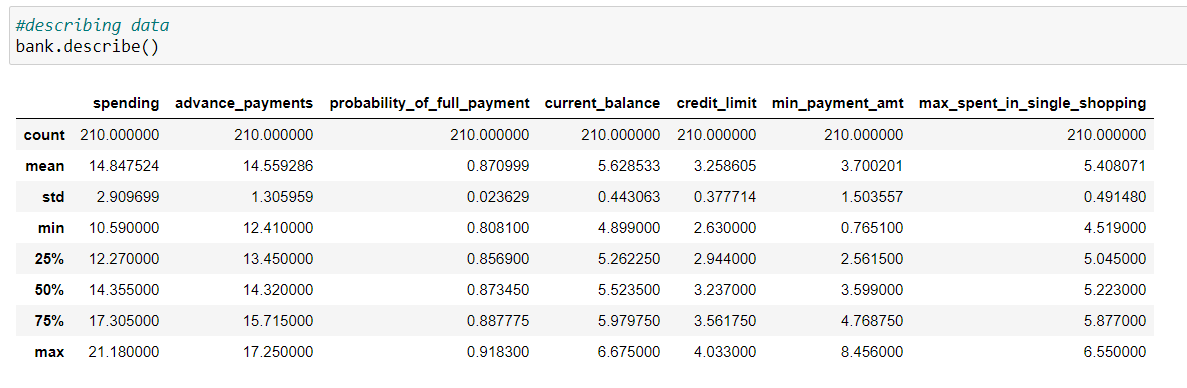
* Checking info of data:-



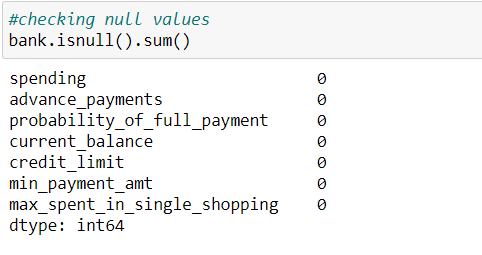
* Checking shape of data:-



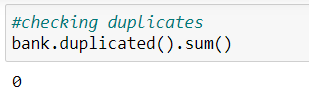
* Describing data:-



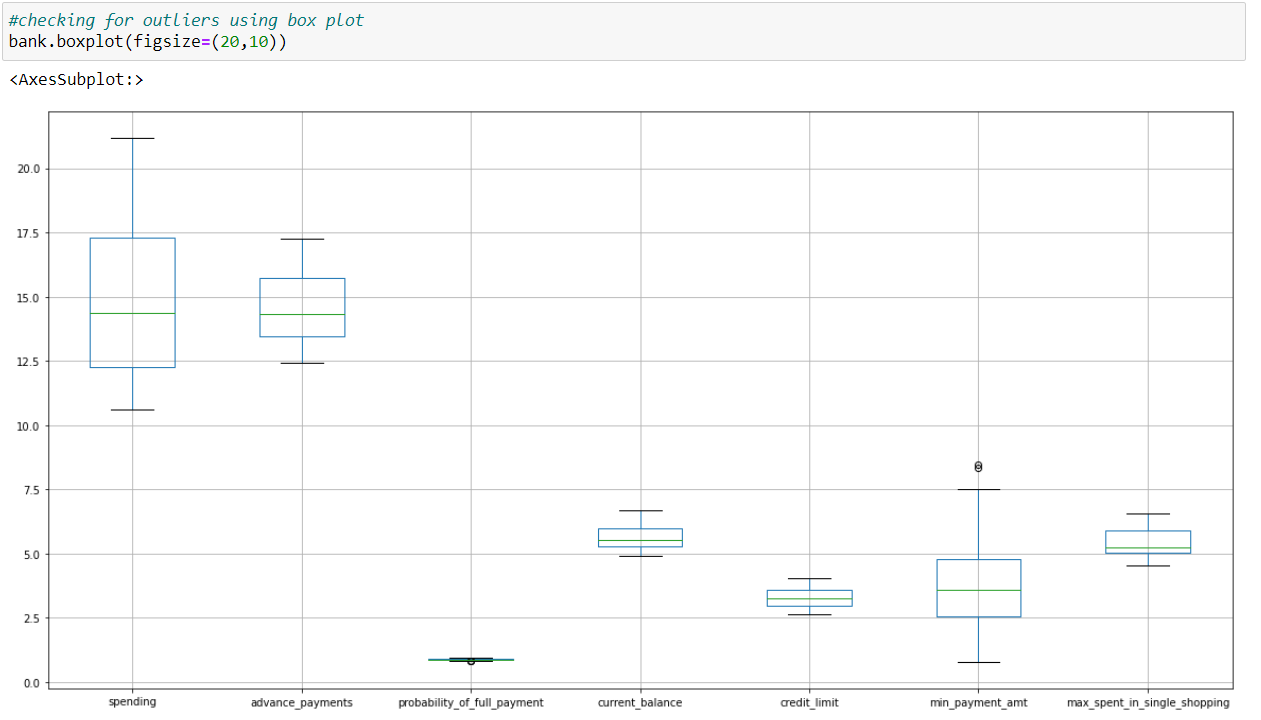
* Checking null values:-



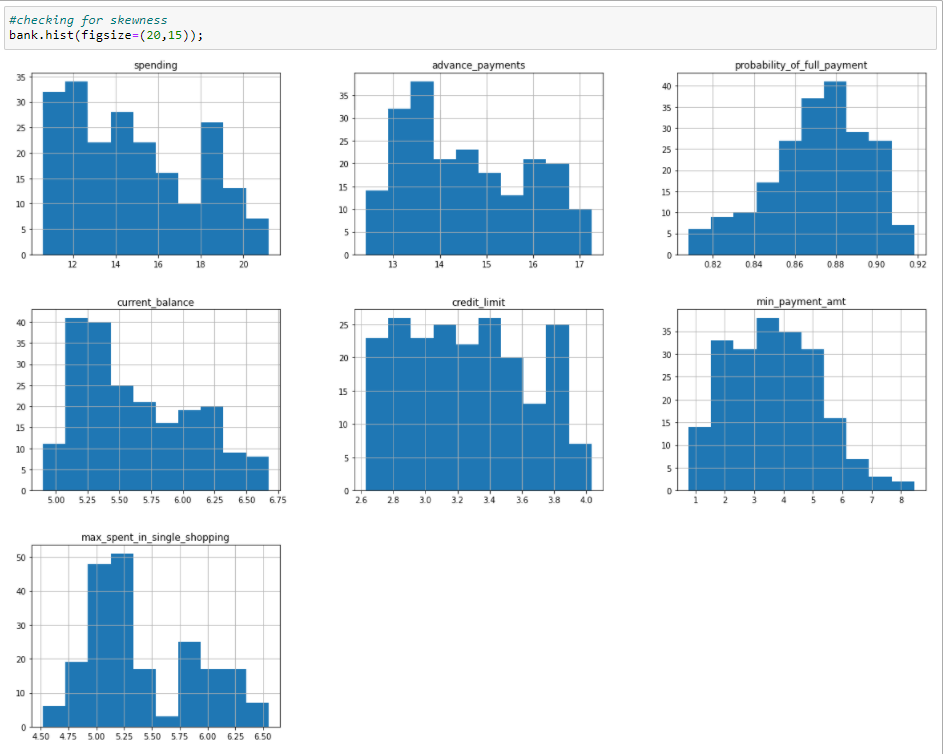
* Checking duplicates:-



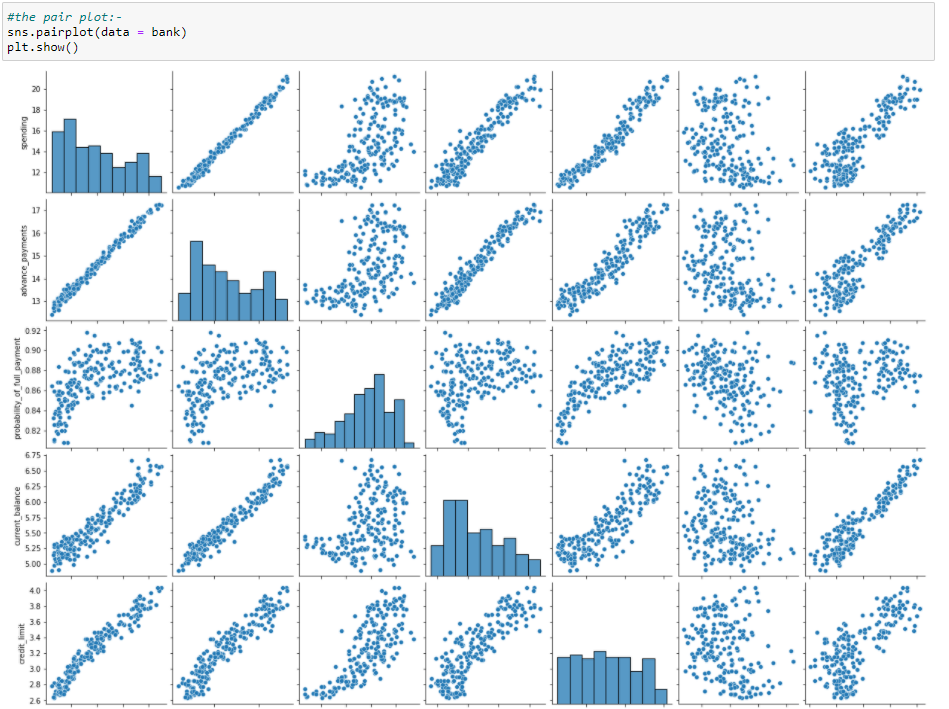
* Checking for outliers using box plot:-



* Checking skewness using histogram:-

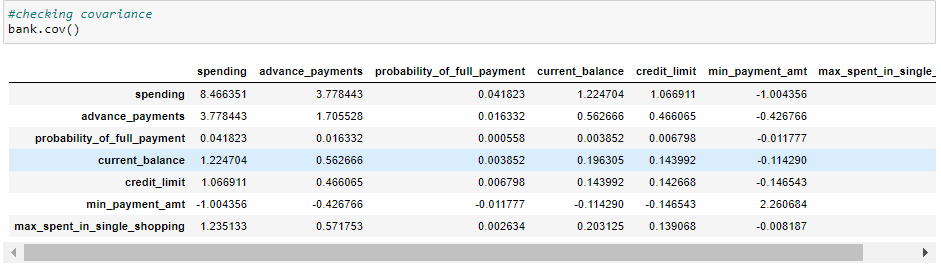


* Pair plots:-

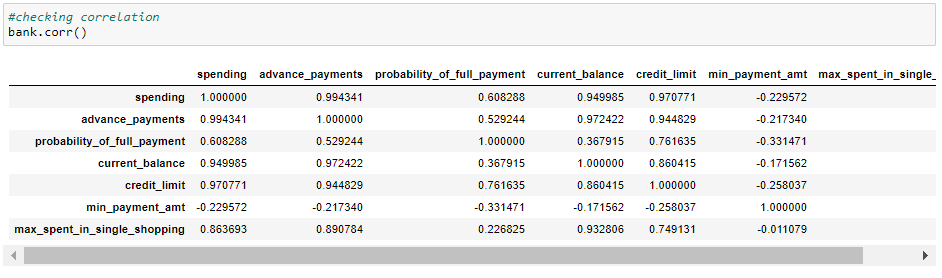




* Checking covariance:-



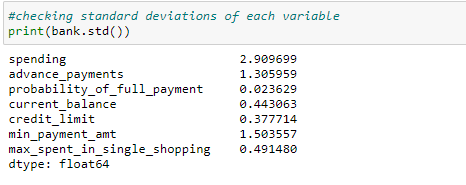
* Checking correlation:-



* Heat map representation of correlation:-



* Checking standard deviation of each variable:-

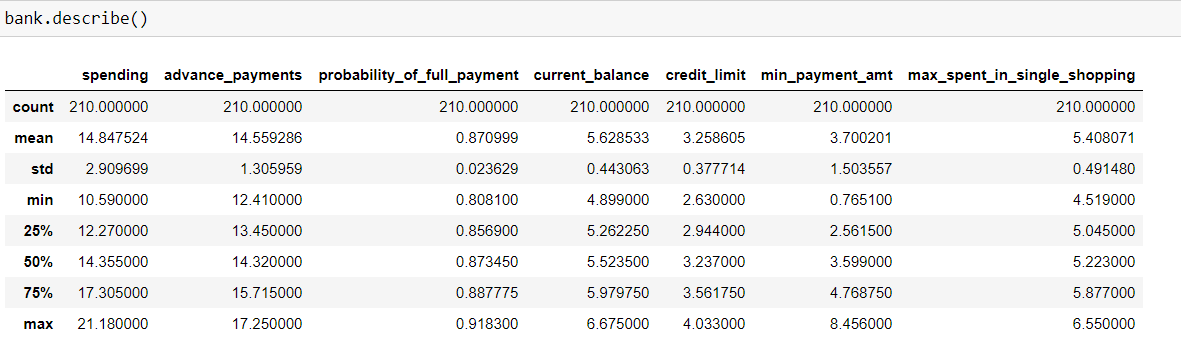


* **INFERENCES FROM THE ABOVE EDA:-**
* While checking the shape we understand that there are 210 rows (observations) and 7 columns (variables).

1. features are:-

* 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)
* We checked info of the data and understand that all the variables are of data type “Float”.
* Data has no null values.
* Data has no duplicate observations
* While describing data we got that all the 7 variables differ in their mean, min, max, standard deviation and IQR values.
* While checking for outlier’s we got that only the variable’s “probability\_of\_full\_payment” and “min\_payment\_amt” have outliers and remaining 5 variables does not have any outliers.
* By plotting histogram for each variable/features we observe that none of the variables are normally distributed. They are either left or right skewed.
* Got different observations for different combinations of variable while checking for covariance and correlation. Few of the variables are highly corelated like spending and credit limit.
* Checked standard deviation for each variable and observe that each variable shows different standard deviation.
  1. Do you think scaling is necessary for clustering in this case? Justify

**Solution:-** We have used the describe function to check on the distribution of data, as shown below:-

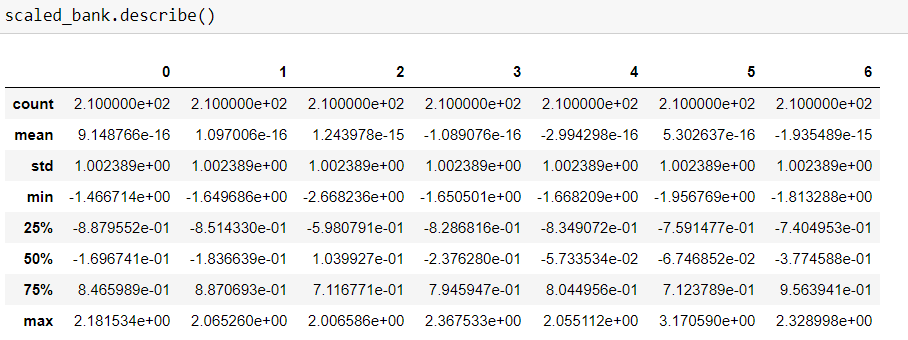


We can see ‘max’, ‘min’ and ‘std’attributes along with all other attributes are showing very different scaling of impact on overall dataset from features, to standardize impact of all features, we need ‘scaling’ is needed.

For this particular problem we are using “standard scalar” function of python to scale the data.

Standard scalar function is an inbuilt package within python under sklearn.preprocessing.

**Describing data post scaling:-**



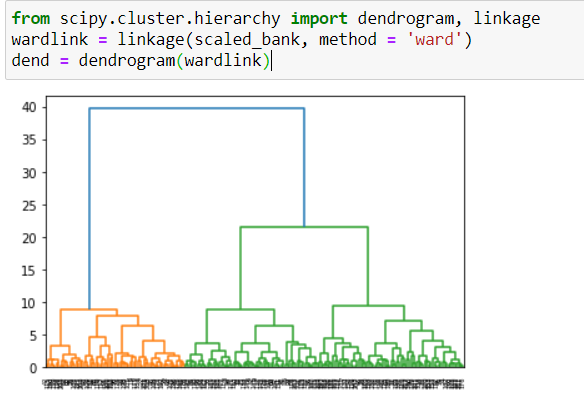
We can observe that post scaling the data is normally distributed across variable and standard deviation along with all other attributes are ranging between -1 and 1.

**How does standard scalar work:-**  StandardScaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. It is also called as Z-score scalling.

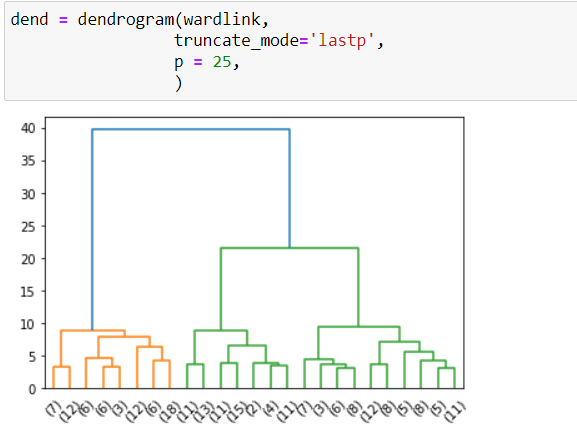
The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1.

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

**Solution:-** Applied hierarchical clustering to the scaled data by using Dendogram and linkage package in python. Where in Dendogram is used for visualization of the cluster formed and linkage is for computing the distance and merging the cluster from n to 1. In this case we have used “Ward” method to form the cluster and when visualized found that there are 2 clusters formed, one in orange colour and other in green colour as shown below:-



Since the above shown cluster is nit messy we have truncated the cluster into last 25 clusters to get a clear picture of the clusters as shown below:-

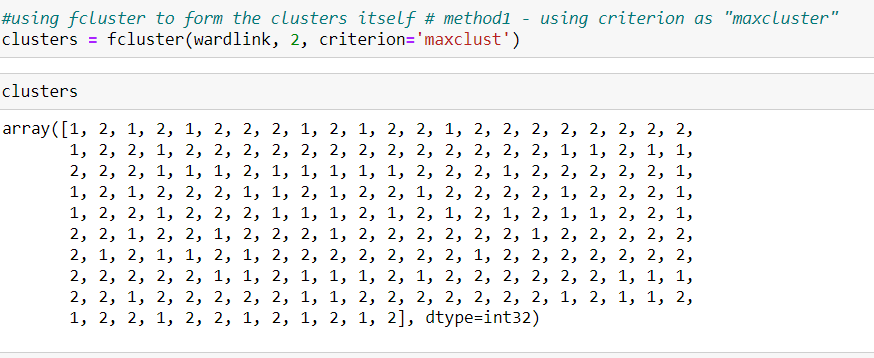


**From the above dendogram visualization we can see that the data has been divided into 2 clusters.**

Now, next step is to label each and every observation with a cluster and to define which observations are into which clusters where are created above. This can be achieved using fcluster function under scipy.cluster.hierarchy

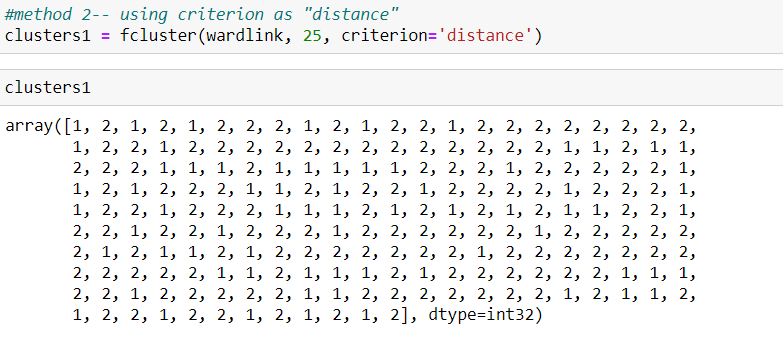
Method 1:- using “maxclust” as criterion

In this we define the number of clusters we are looking for.



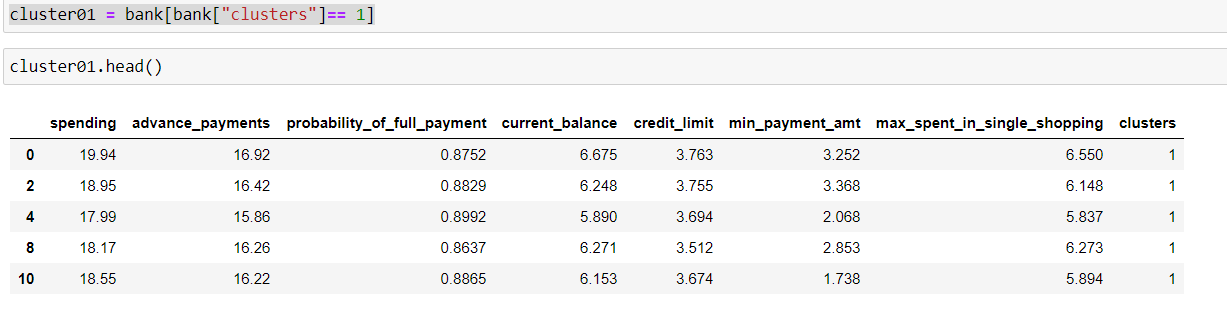
Method 2:- using “distance” criterion

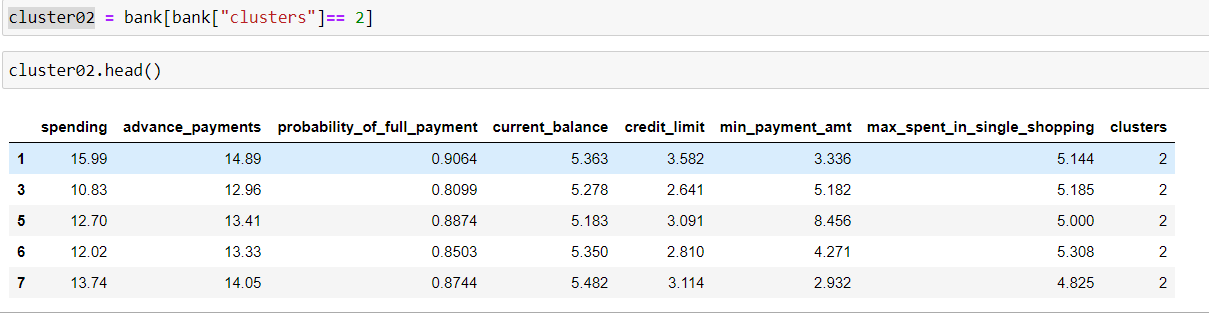
In this method we use the distance for which we need to create the clusters.



With the help of dendogram we have concluded that there are 2 number of clusters formed and we have also restricted fcluster function also till 2 number of clusters.

We can divide data into 2 clusters as “cluster01” with all observations from 1st cluster and “cluster02” with all observations from 2nd cluster.

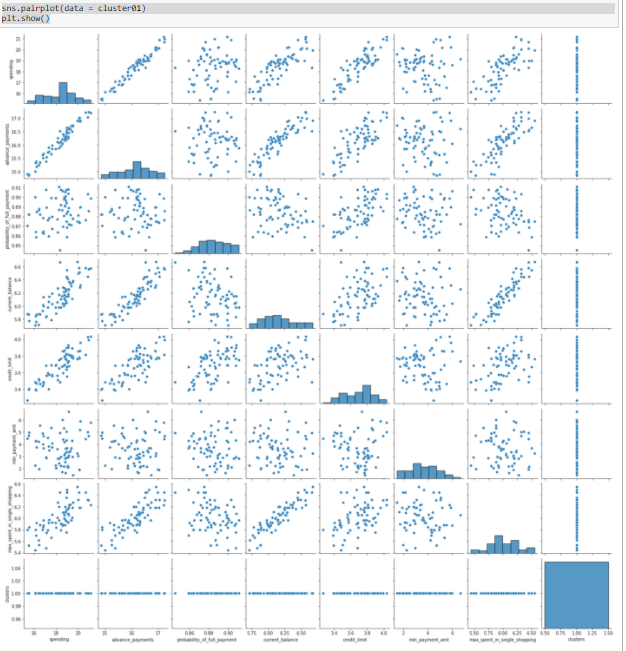


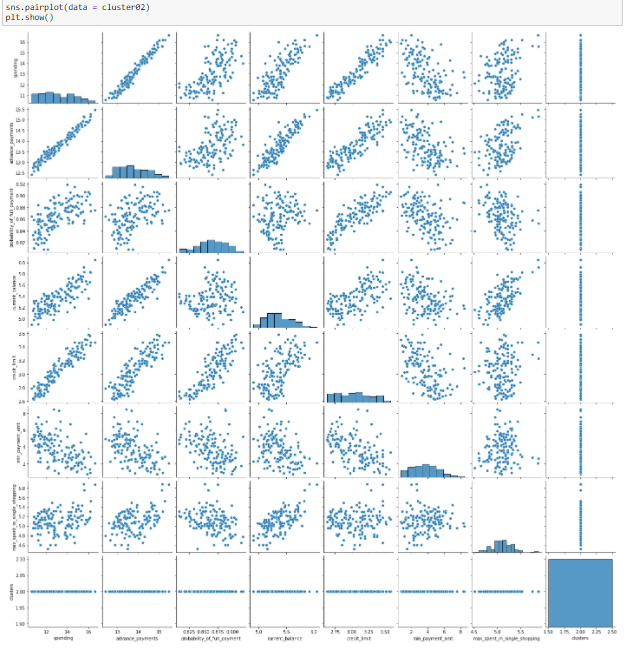


Below are observations from two clusters:-

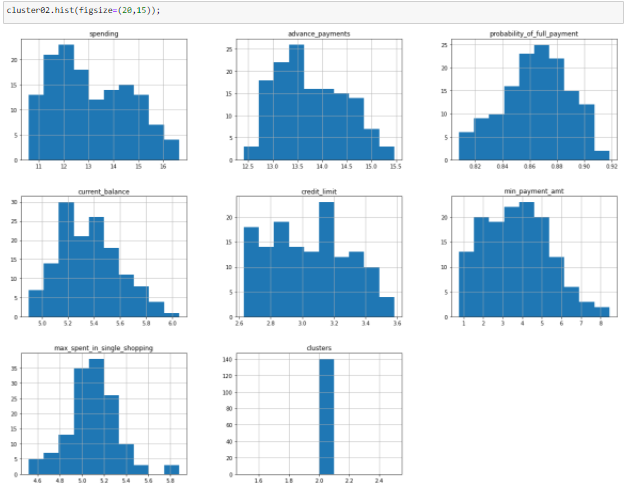
* 1st cluster have 70 observations and 2nd cluster have 140 observations
* In 1st cluster the customers are good with advance payments as compared with 2nd cluster.

Used below plots to visualize these two different cluster:-









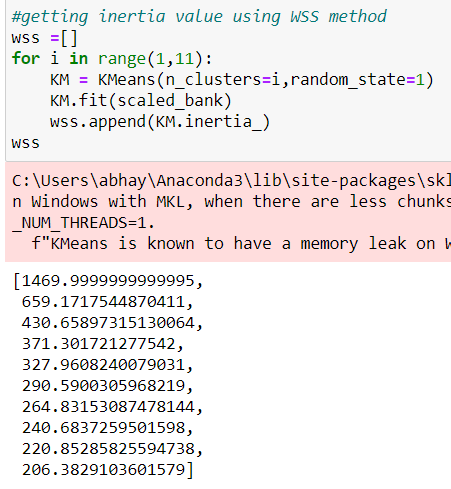
* 1. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

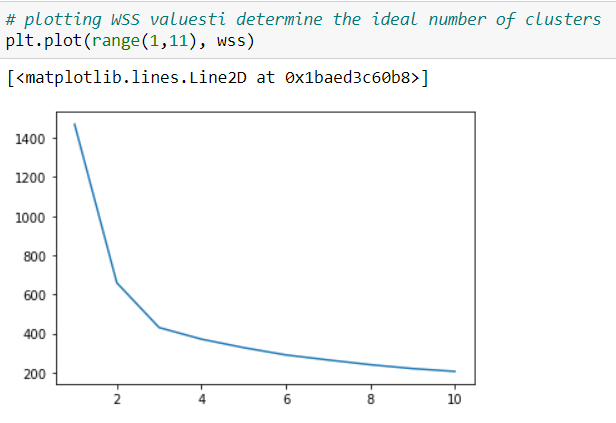
**Solution:-**

applied K means clustering and while checking for inertia values found that the drop in inertia value from 1 to 2 number of clusters is remarkable however, when we move from 2 number of clusters to 3 number of clusters the drop in inertia value is not remarkable as compared. And inertia value from 3 number of clusters to 4th cluster is ever low.

Using WSS we got the below inertia values from 1 cluster till 11th cluster:-

Plotting these WSS values in a plot to under the optimal number of clusters as shown below:-





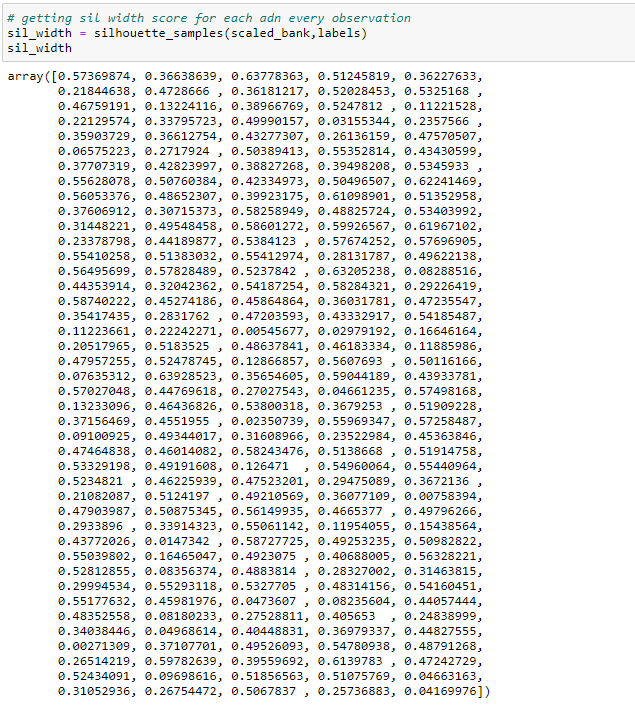
* Looking at the above WSS score we can conclude that the 3 would be the optimal number of clusters as we can see remarkable difference in WSS score.

Lets validate the optimal number of clusters using silhouette score:-

Lets now check silhouette score with optimal number of clusters as 3. We find that the silhouette score is 0.40072 and we understand that the score is decent and indicates good clustering.



Lets check the sil width score for each observation:-

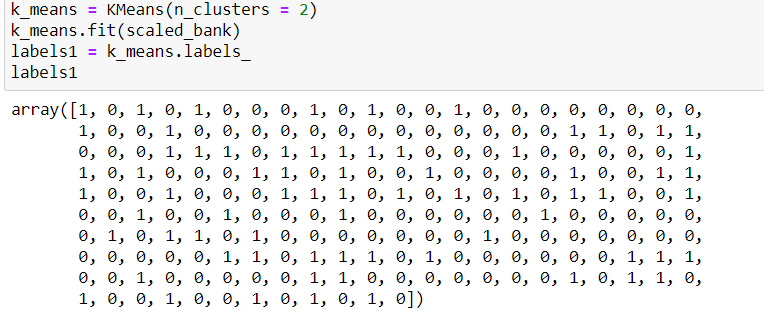


Lets check the minimum value of the sil width score:-

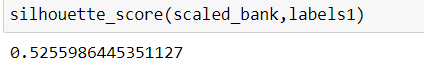


Here by we can conclude that 3 is the optimal number of clusters as for few of the observations sil width score is in positive.

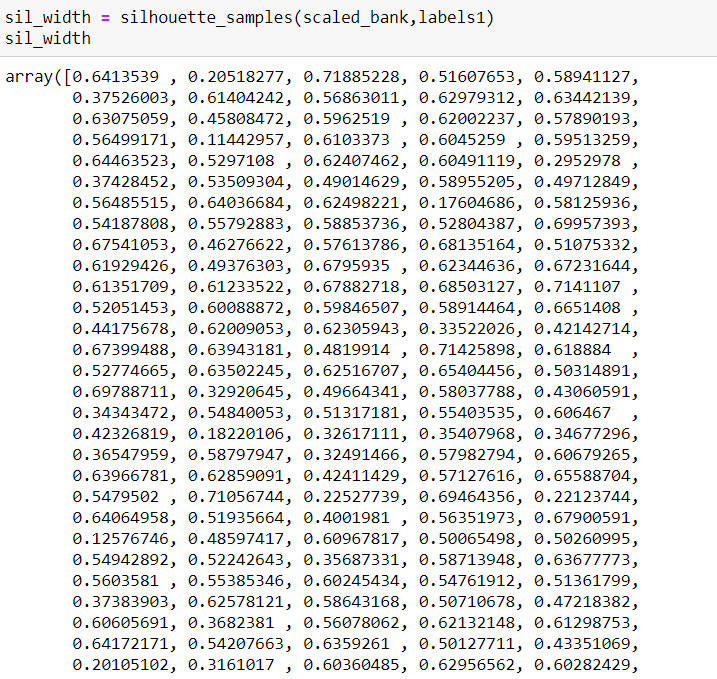
Lets try with optimal cluster as 2:-



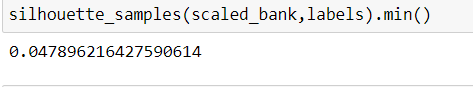
Lets now check the silhouette score with 2 as the optimal number of clusters:-



The value shown above is more than 0.5 and indicates that the clustering is properly done. Lets now check the sil width score of each observation:-

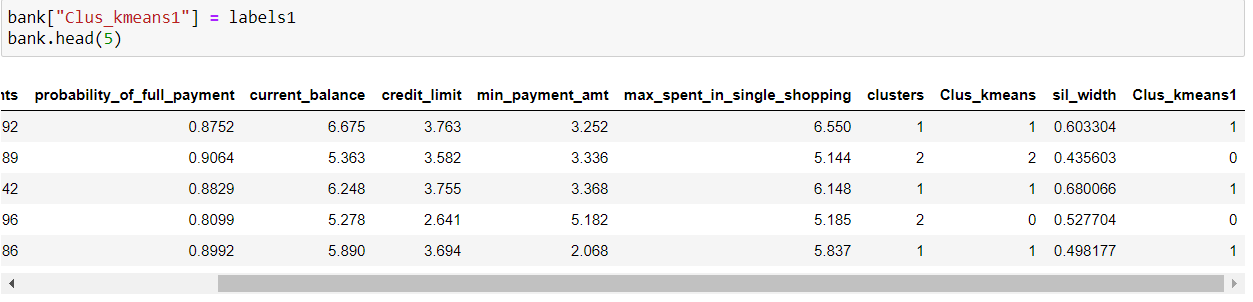


Lets check the min value of this new sil width score with number of clusters as 2:-

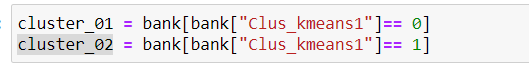


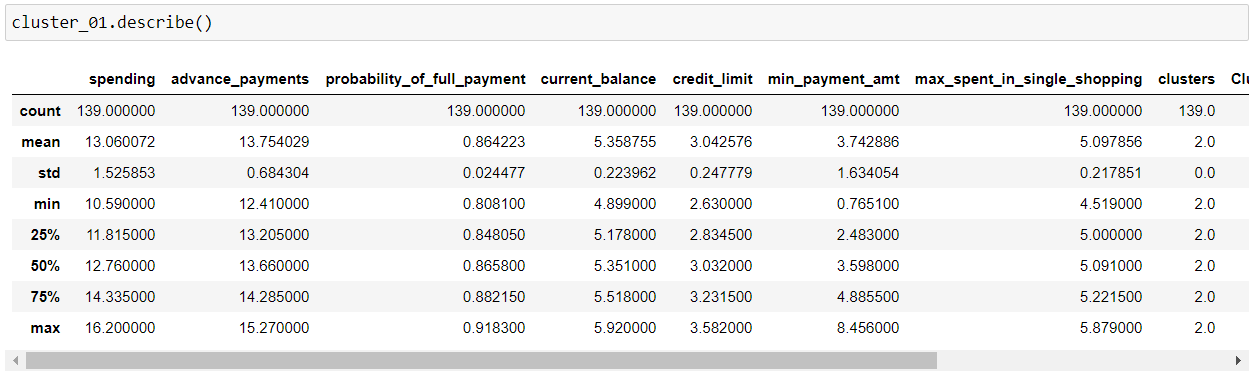
The min value of the sil width score is positive.

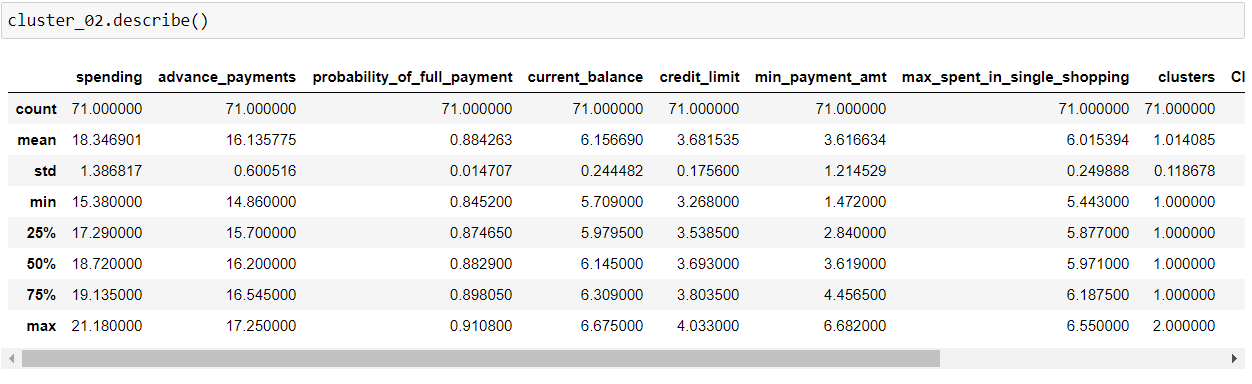
Appending cluster labels to the original data frame:-



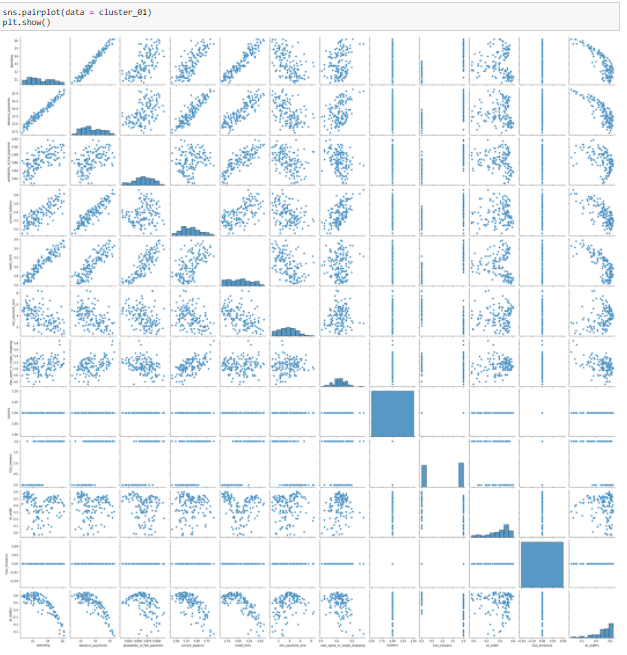
Dividing the data into 2 clusters:-



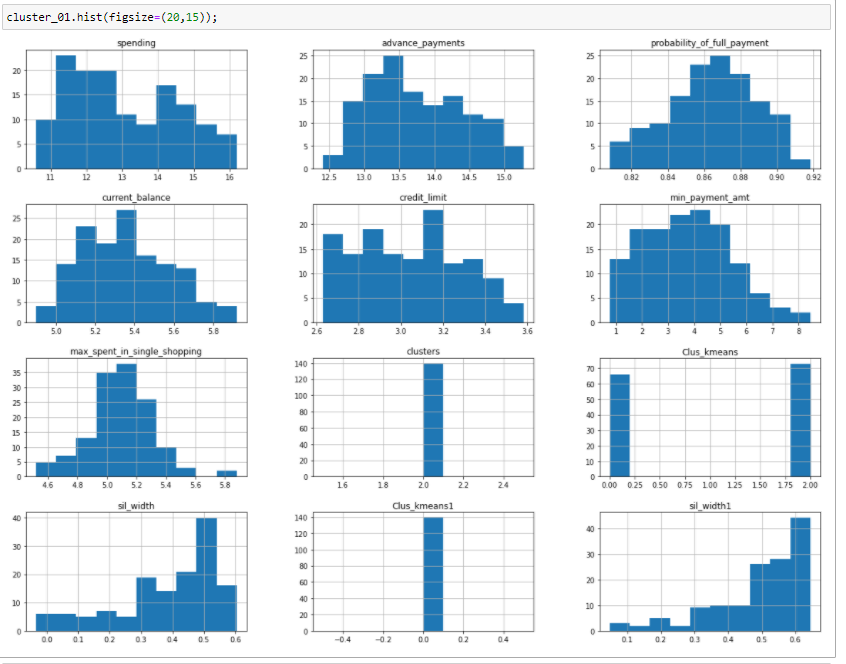


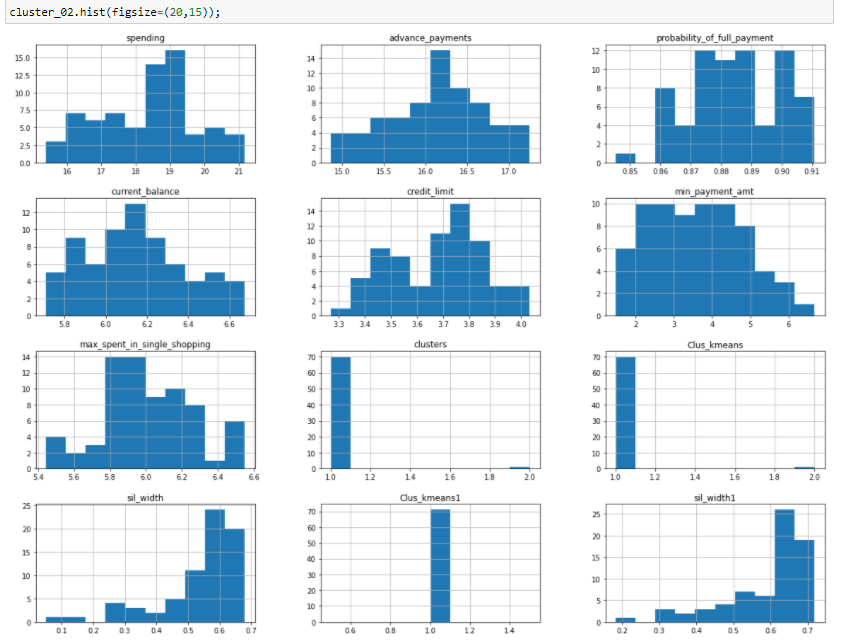


Visualizing customer segmented data:-





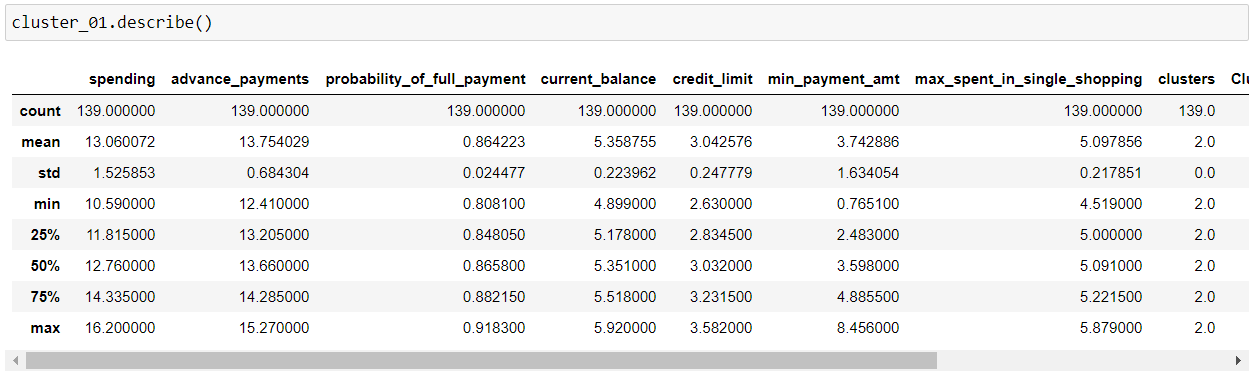


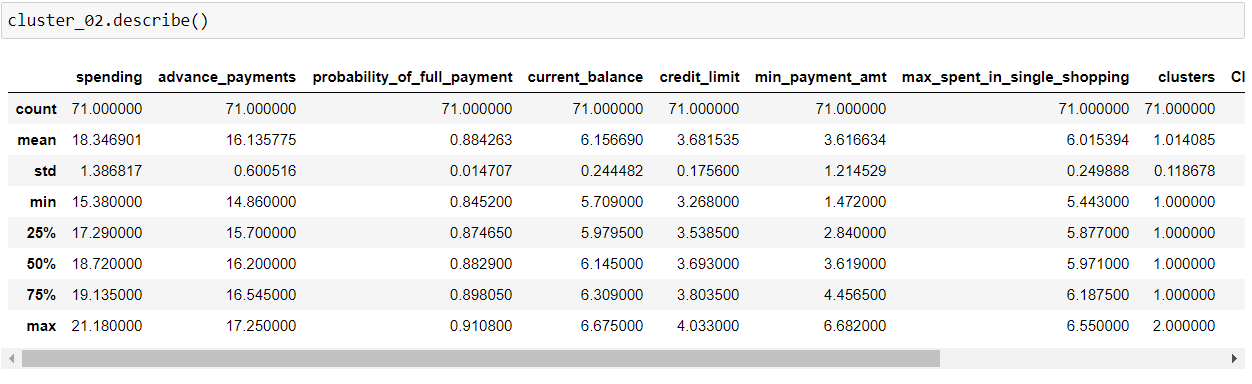


* 1. Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

**Solution:-**

While describing the two clusters forms as shown below:-





**CLUSTER PROFILING: -**

* Cluster 1 contains 139 customers and 2nd cluster contains 71 customers.
* Comparing the means of the customers in both the clusters we can see that the customers on 2nd cluster is high on spending as well as on payments.
* Probability of making full payment is higher for customers in 2nd cluster.
* The online spending is more for the customers in 2nd cluster.
* The probability of min due amount payment is more of customers in 1st cluster.
* We can conclude that the customers in 2nd cluster are the customers with high spending and with timely payments and chances of missing on the payment is almost NIL.
* For 1st cluster we are mediocre in spending but may default in making the complete or advance payment.

**RECOMMENDATIONS:-**

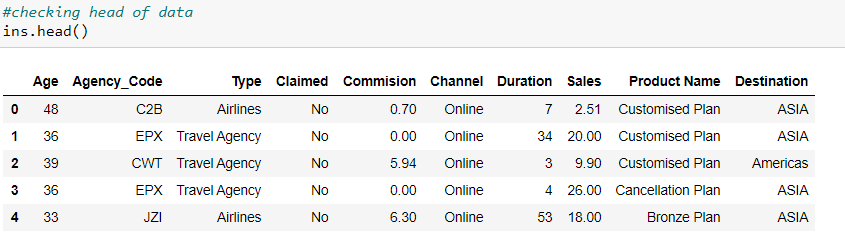
* For customers in 1st cluster business can give rewards point on making full payment or advance payment.
* For customers in 2nd cluster business can roll out seasonal offers to increase their spending on credit card along with more reward points.
* For customers in 1st cluster may give an option to customer to increase in credit limit after making full payment of advance payment.
* For customers in 1st cluster can provided with offers on credit card for family members upon making full payment for few months.
* For customers in 2nd cluster business can provide discount vouchers over min amount of spending’s to increase their spending on the credit card.

**Problem 2: CART-RF-ANN**

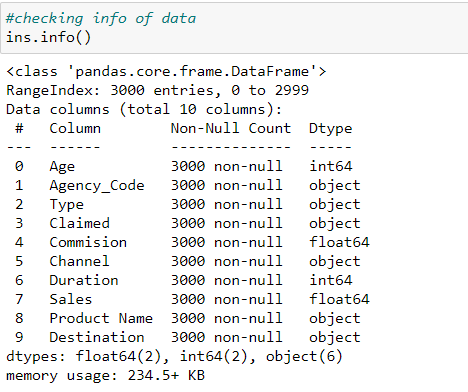
An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models' performances in train and test sets.

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

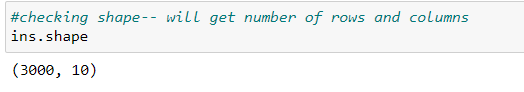
**Solution:-** **Solution:-** lets check head of data:-



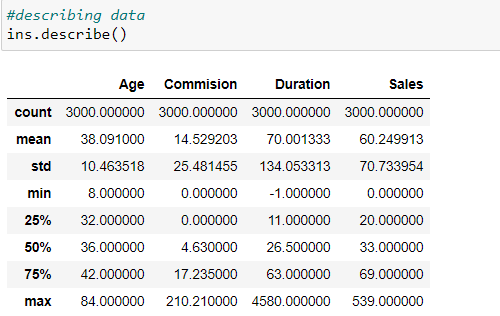
* Checking info of data:-



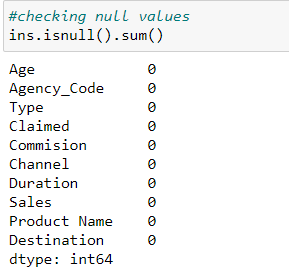
* Checking shape of data:-



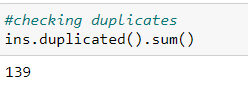
* Describing data:-



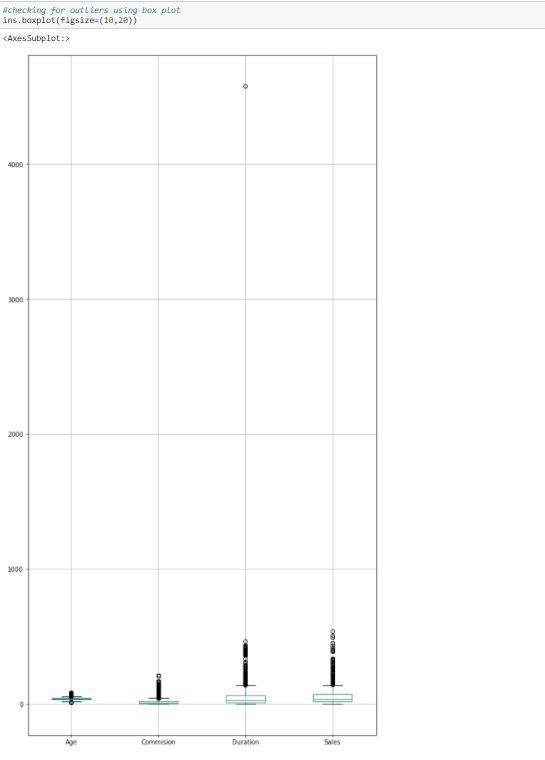
* Checking null values:-



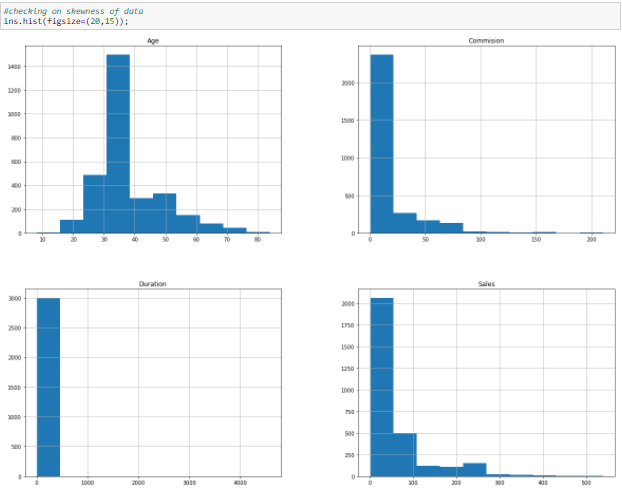
* Checking duplicates:-



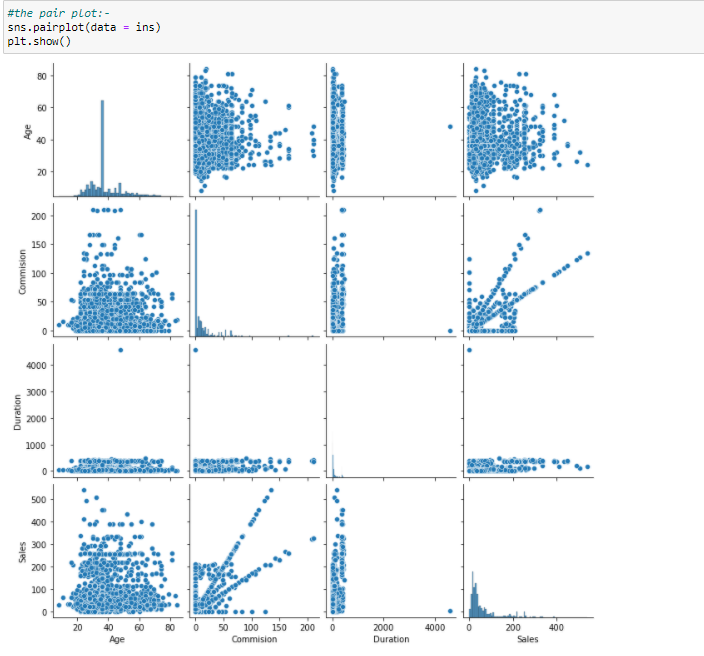
* Checking for outliers using box plot:-



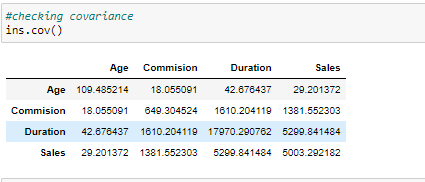
* Checking skewness using histogram:-



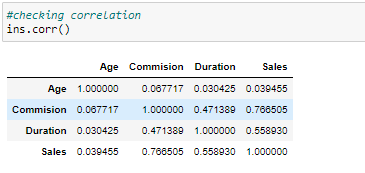
* Pair plots:-



* Checking covariance:-



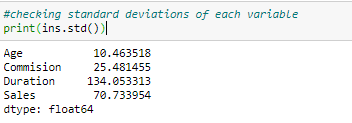
* Checking correlation:-



* Heat map representation of correlation:-



* Checking standard deviation of each variable:-



* **INFERENCES FROM THE ABOVE EDA:-**
* While checking the shape we understand that there are 3000 rows (observations) and 10 columns (variables).

1. features are:-

1. Target: Claim Status (Claimed) – dependent variable

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)

7. Destination of the tour (Destination)

8. Amount of sales of tour insurance policies (Sales)

9. The commission received for tour insurance firm (Commission)

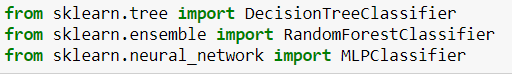
10. Age of insured (Age)

* We checked info of the data and found that the columns “agency code”, “type”, “claimed”,”channel”,”product name” and “destination” are of data type object and rest other 4 features are of data type float and integer.
* Data has no null values.
* Data has 139 duplicate observations
* While describing data we got that all the 4 integer and float variables differ in their mean, min, max, standard deviation and IQR values.
* While checking for outlier’s we got that all the 4 integer and float variables has got outliers in the data.
* By plotting histogram for each variable/features we observe that none of the variables are normally distributed. They are either left or right skewed.
* Got different observations for different combinations of variable while checking for covariance and correlation. None of the variables are highly correlated with each other.
* Checked standard deviation for each variable and observe that each variable shows different standard deviation.

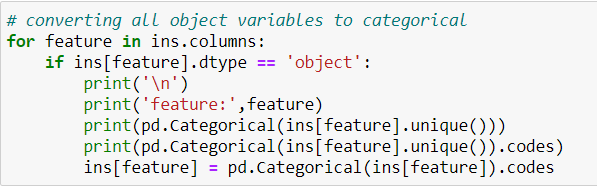
**2.2** Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network

**Solution:-**

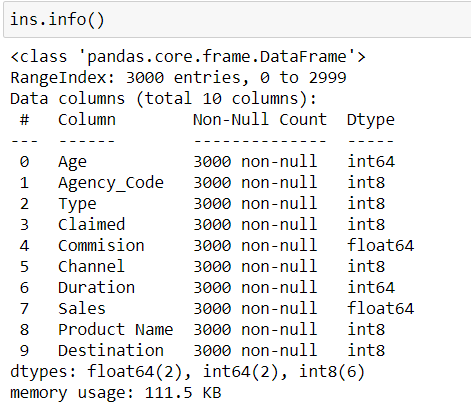
* Loading necessary packages for building classification models

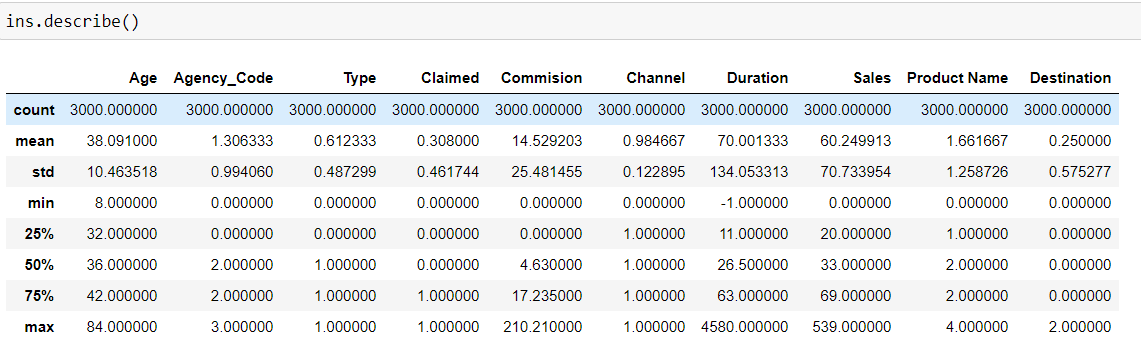


* Converting all object data type variables into categorical

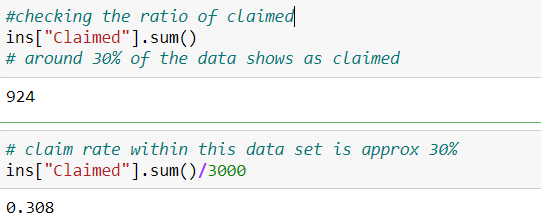


* Checking info and description of data post conversion of object variables into categorical

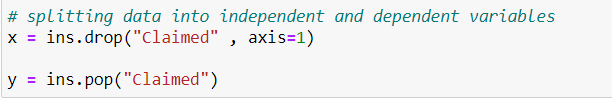




* Checking the ratio of claimed one’s from the data set and it shows that around 30% of the customers have claimed insurance from the given data set.



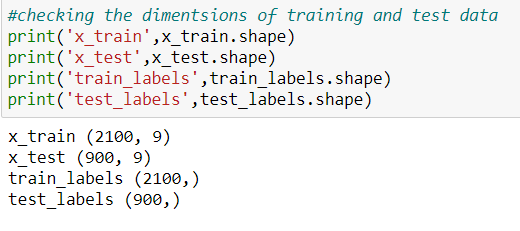
* Dividing data into dependent and independent variables. Where “x” contains all independent variable’s and “y” have only dependent variable.



* Splitting the data set into 70% training and 30% testing as per industry standards or as per accepted market practices with using “random state” function as “1”.

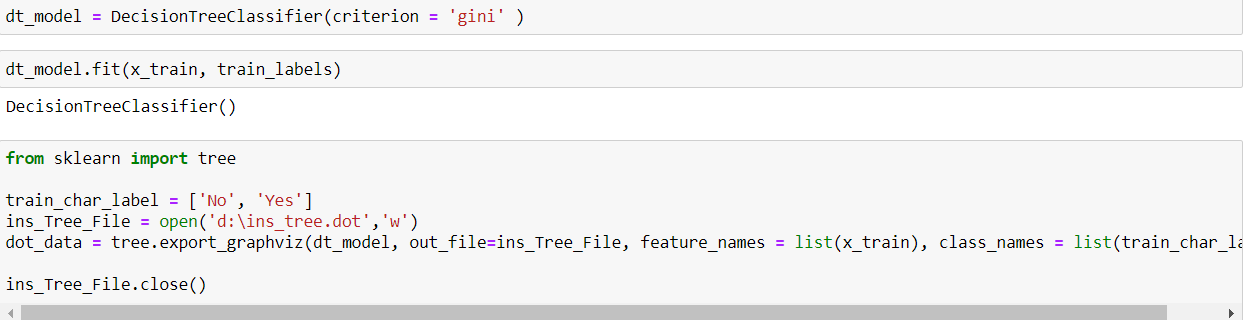


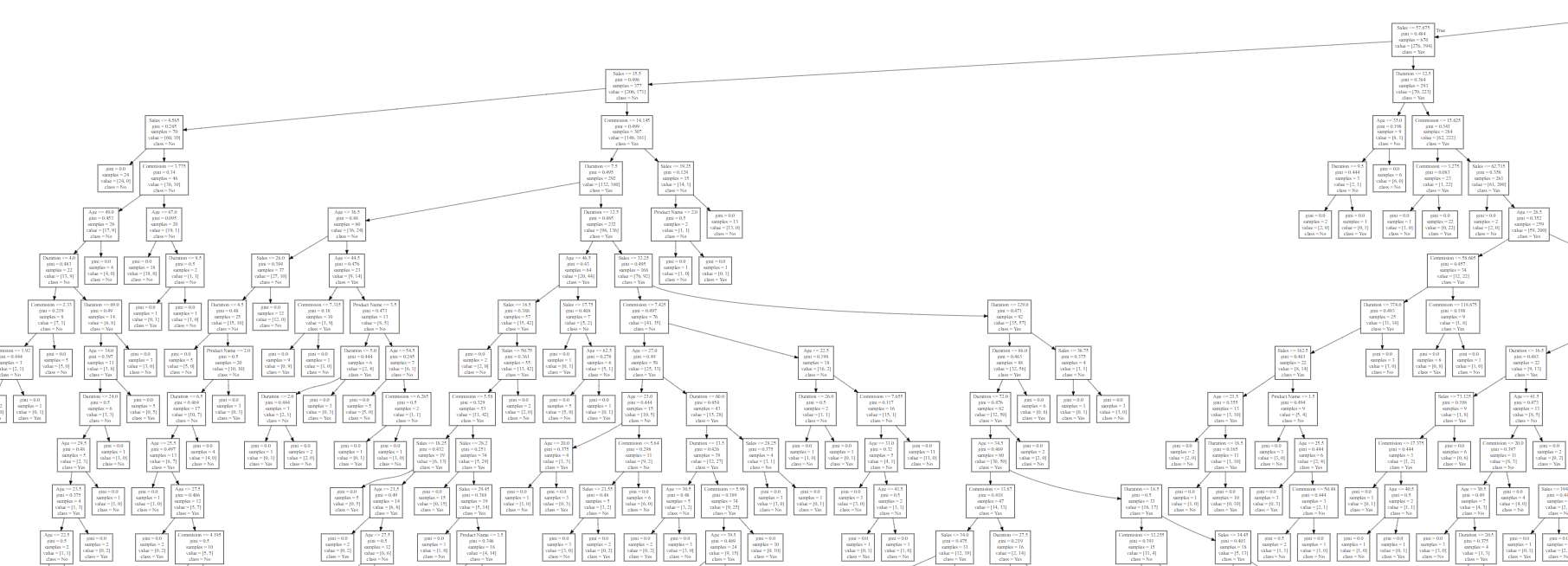
* The test and training data set looks like shown below along with their test and train labels: -

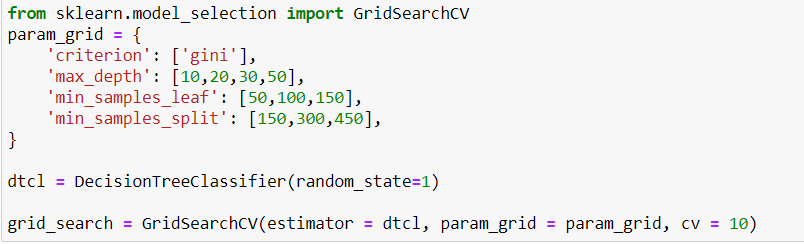


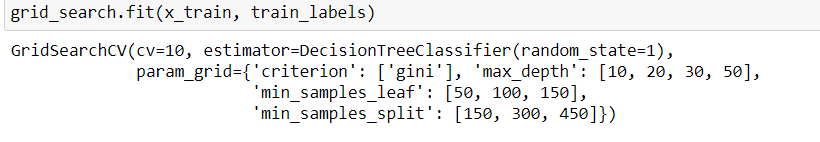
* **BULIDING A DECISION TREE CLASSIFIER MODEL:-**

Firstly I build the model with gini as criterion without pruning the data and visualized the data using webgraphwiz.com . As shown below:-

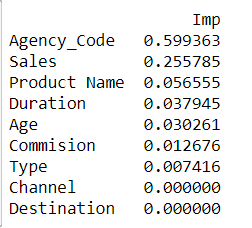




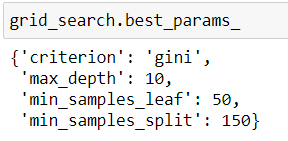
* Since the decision tree is over grown now we need to prune the data to get reasonable number of branches and leaf nodes. Which can be achieved by using “gridsearchCV” function under sklearn.model\_selection.
* 
* Min sample leaf = 1-3% of the entire data and min sample split = 3 times of min sample leaf.
* Fitting grid\_search into training data set:-



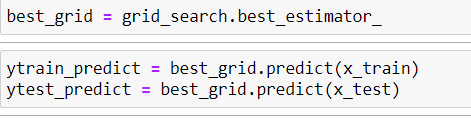
* Checking feature importance for CART model:-



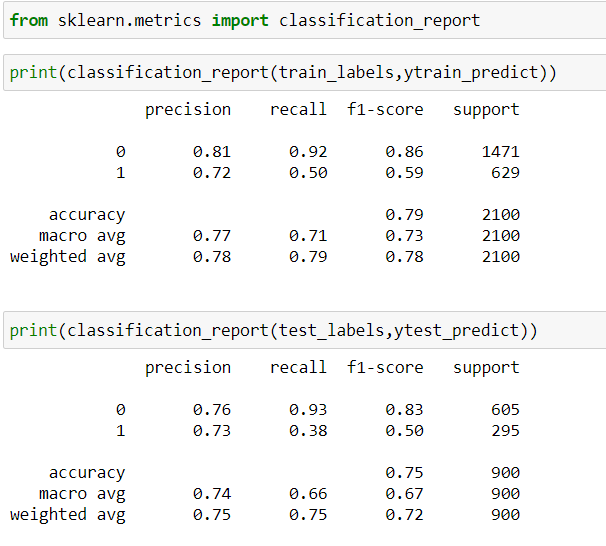
* Getting best parameters:-



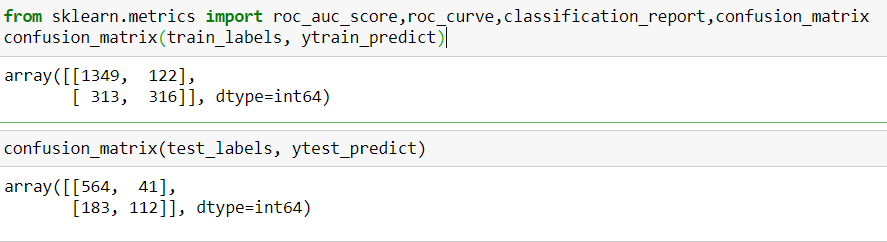
* Getting best estimators and using these best estimators for prediction



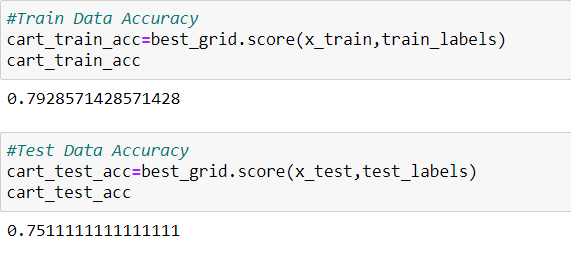
* Classification report for prediction for train and test data sets



* Creating confusion matrix for prediction for train and test data sets

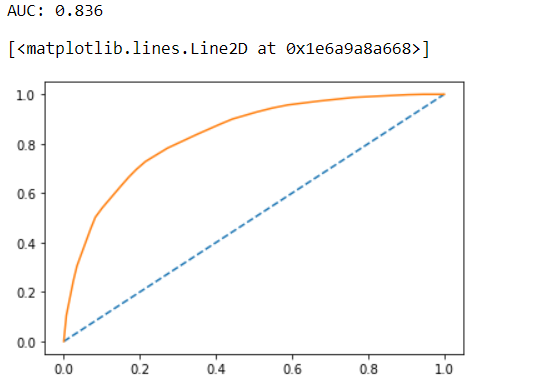


* Checking accuracy for train and test data sets

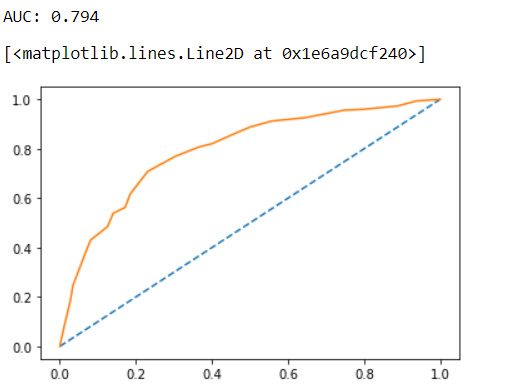


* Getting AUD value and plotting ROC curve for training and testing data sets

For training data prediction:-



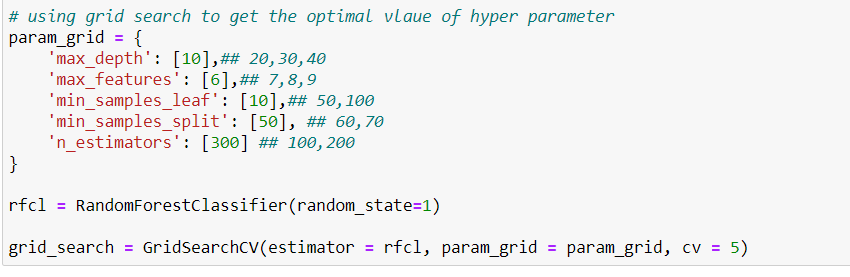
For testing data prediction:-



**By building this CART (decision tree) model we got 79% accuracy for training data set and 75% accuracy for test data set.**

* **BULIDING A RANDOM FORREST MODEL:-**

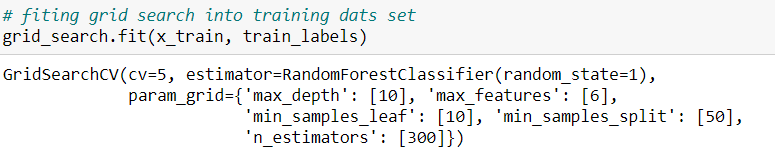
We can use almost the same hyper parameter in grid search CV function as used in decision tree.



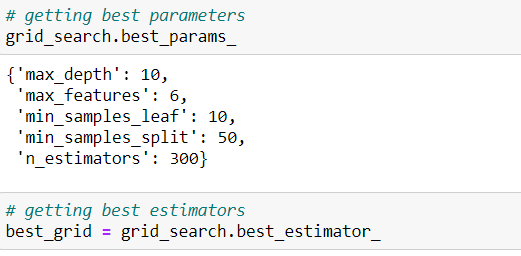
I have added one more additional hyper parameter which is “n\_estimators” which denotes the number of tree we need in the forest we are creating.

Also we need to consider that higher the number of parameter’s and higher the CV value, the system will take longer to give the output. i.e.; the system run time will be more.

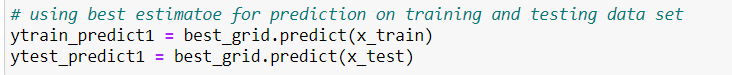
* Fitting grid search into training data set.



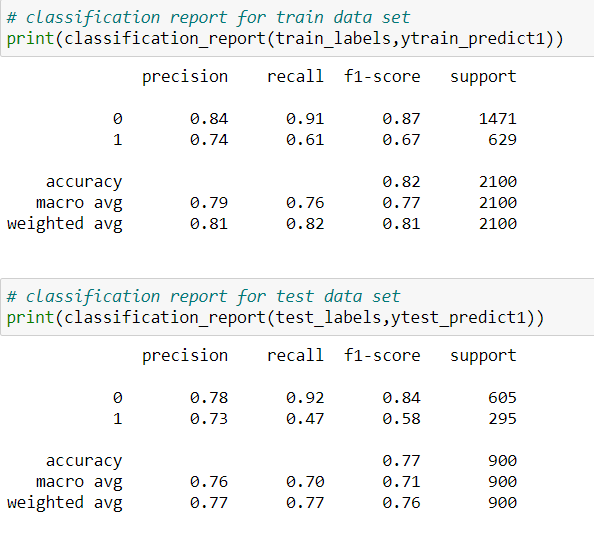
* Getting best parameter and best estimator



* Using best estimators to predict training and testing data set



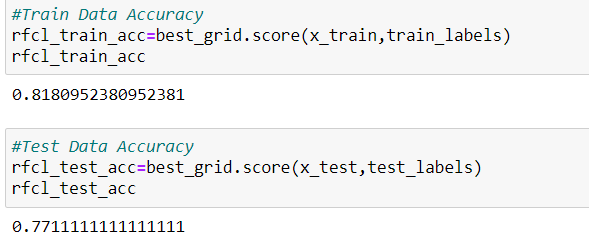
* Creating classification report for training and testing data sets



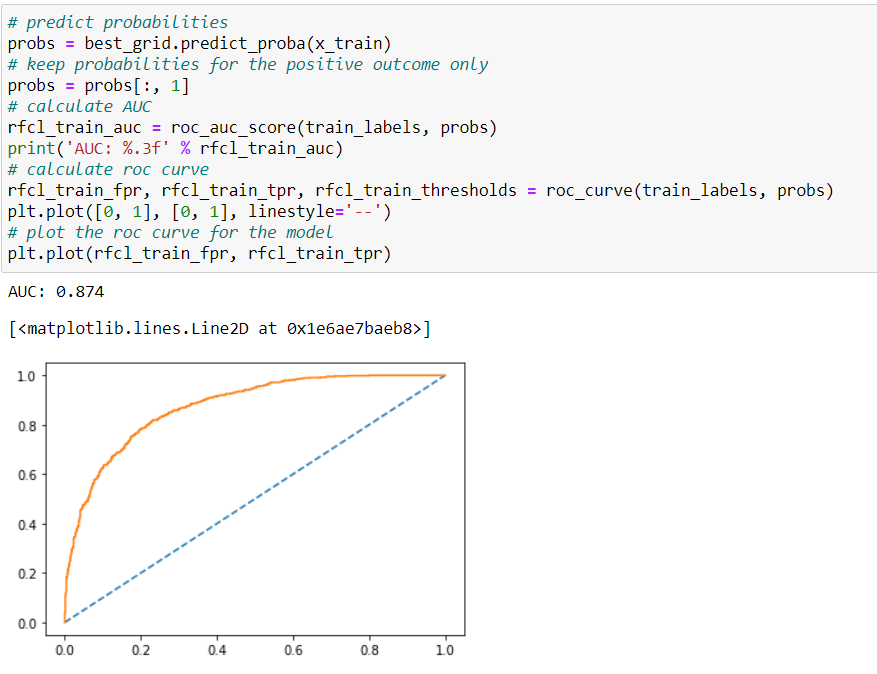
* Creating confusion matrix for training and testing data set



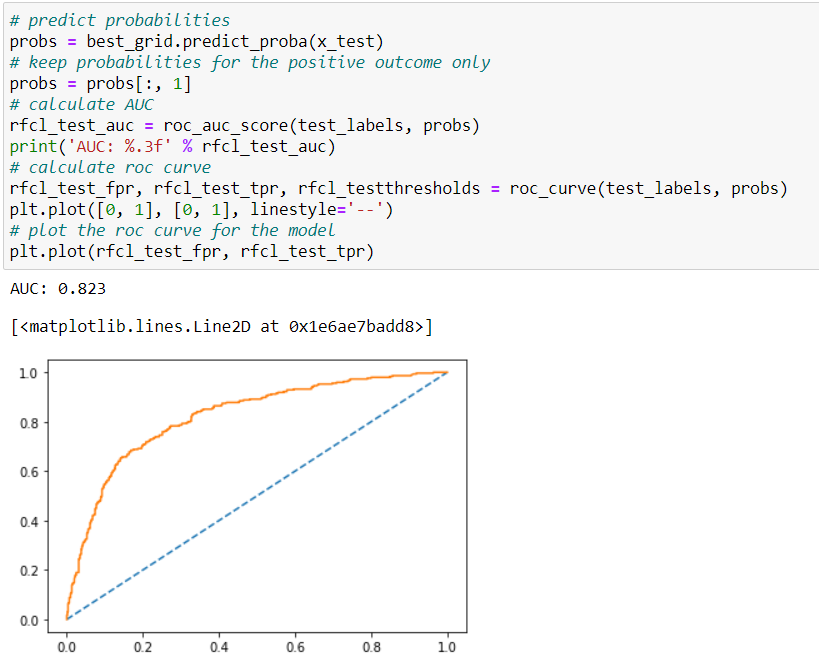
* Checking accuracy for training and testing dataset



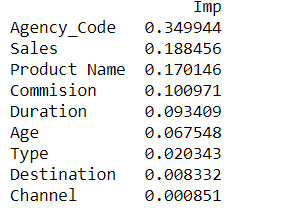
* AUC score and ROC curve for training data set



* AUC score and ROC curve for testing data set

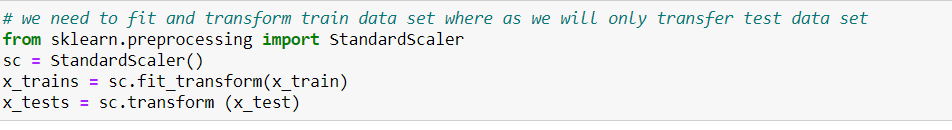


* Checking on the variable importance fir random forest model

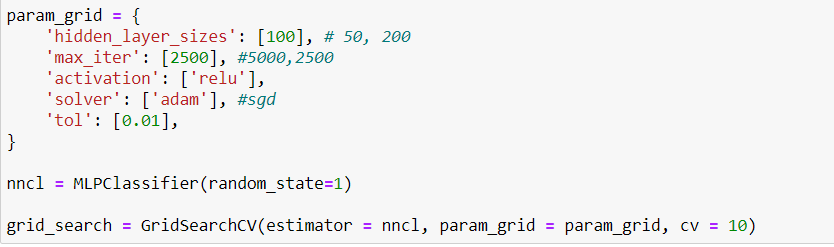


**BUILDING MLP CLASSIFIER MODEL**

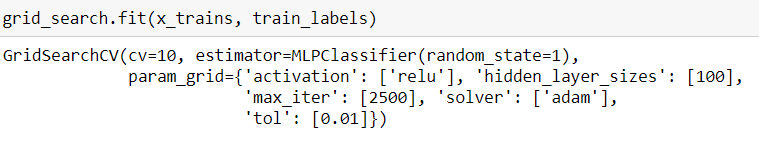
* The only requirement to build an MLP classifier is that the data needs to be scaled. In this case we are scaling the data using standard scaler function available under sklearn.preprocessing.
* Scaling the data



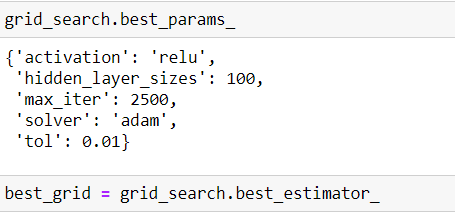
* We have used fit\_transform function for training data set but used only transform function for testing data set. So that the test data set wont distribute same as training data set.
* Using grid search CV function to get the optimal values of hyper parameter



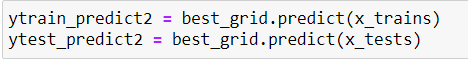
* Hidden layer: number of neurons in hidden layer
* Max\_iter:- number of iterations
* Activation:- activation function used
* Solver:- solver method used
* Tol:- the tolerance level
* Fitting gridsearch into training data set



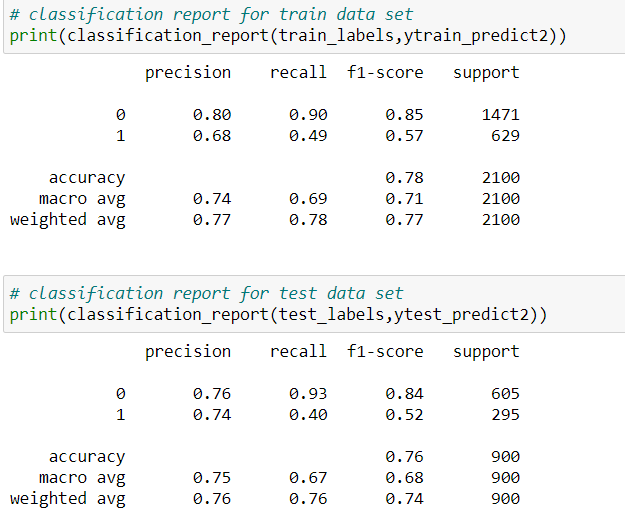
* Getting best parameters and best estimators:-



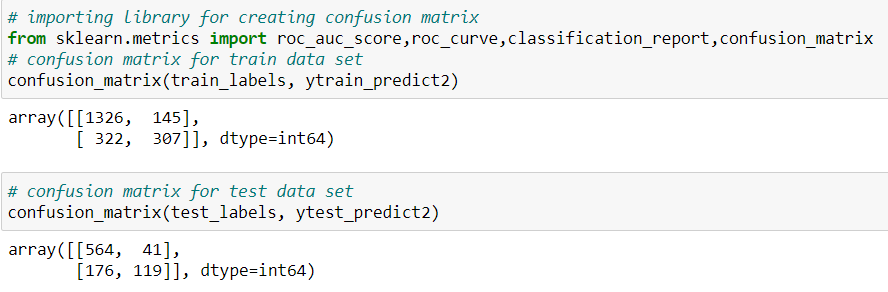
* Predicting training and test data set using the best estimators



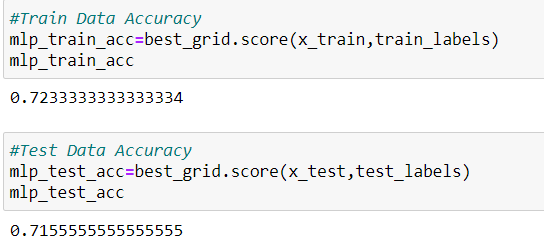
* Getting classification report for training and testing data set



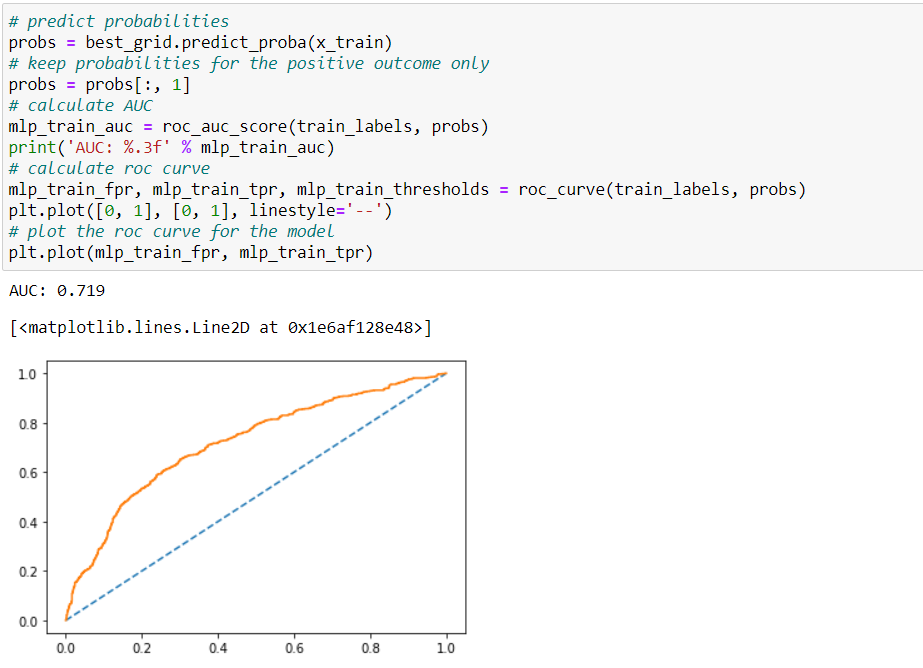
* Getting confusion matrix for training and testing datasets:-



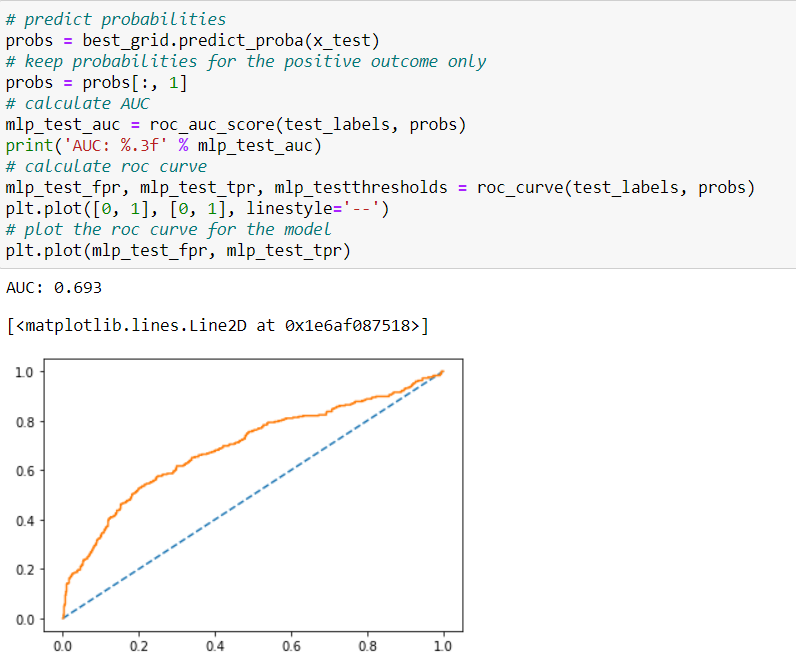
* Getting accuracy for training and testing data sets



* Getting AUC score and ROC curve for training data set:-



* Getting AUD score and ROC curve for testing data set:-

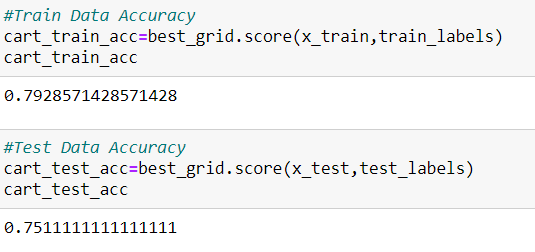


* For MLP classifier model there is no function available for feature importance.

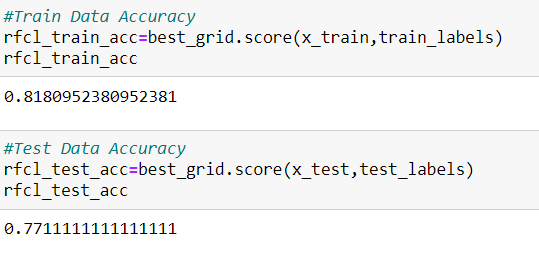
**2.3** Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

**Solution:-**

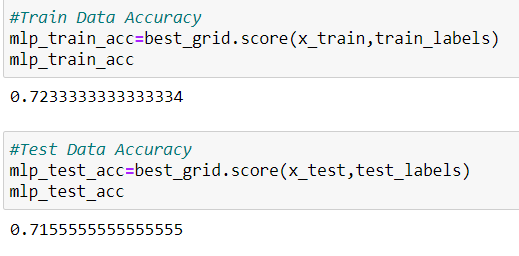
Accuracy for CART model:-



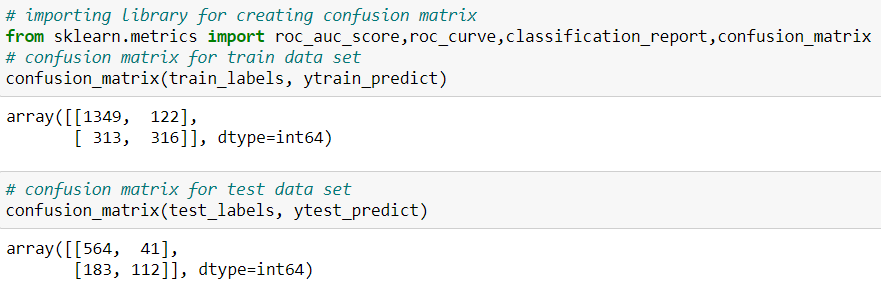
Accuracy for random forest model



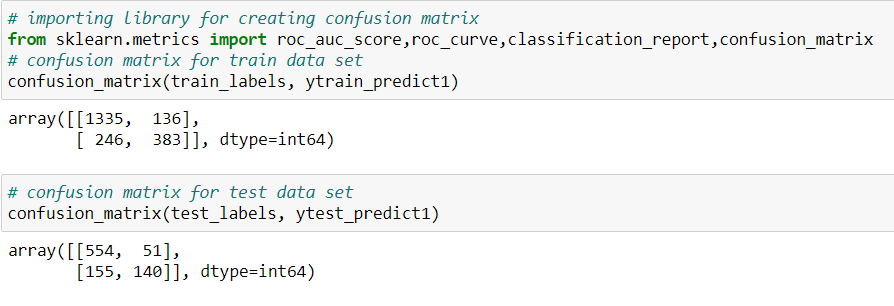
Accuracy for MLP model



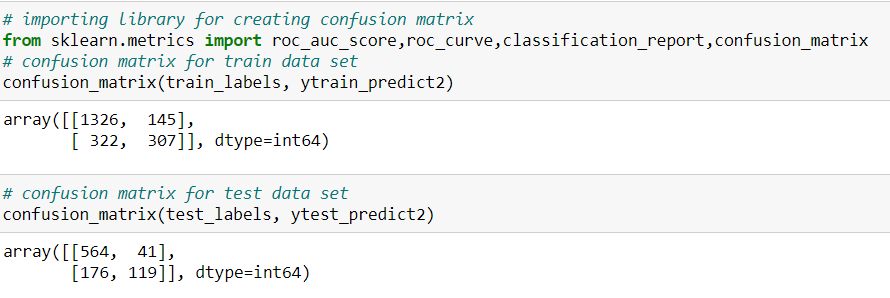
Confusion matrix for CART model:-



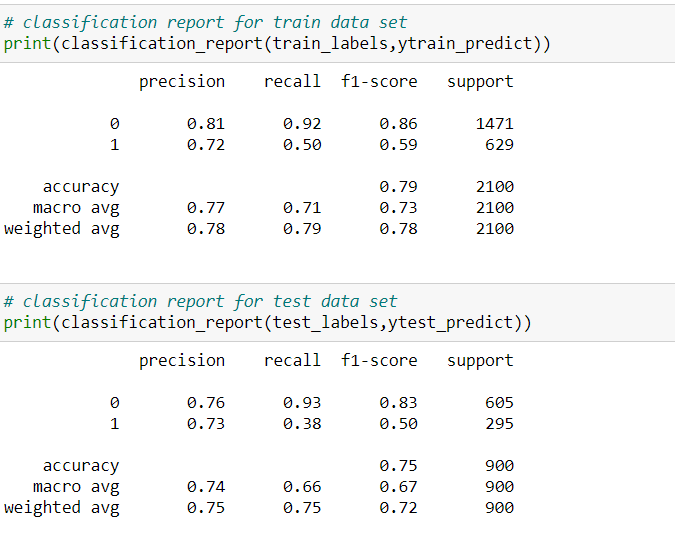
Confusion matrix for random forest model



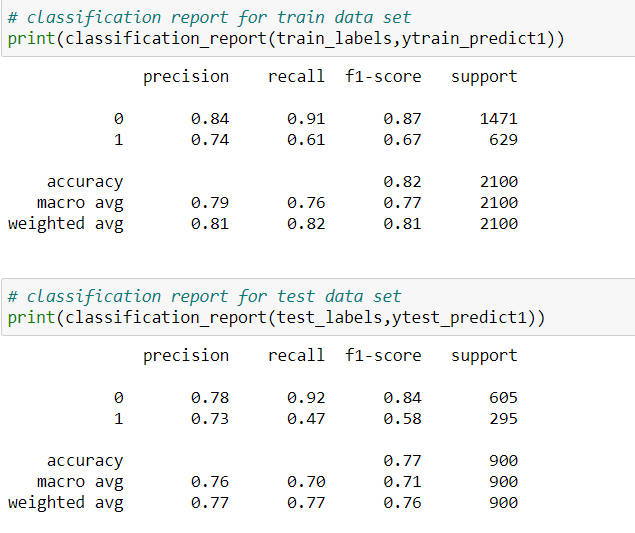
Confusion matrix for MLP model



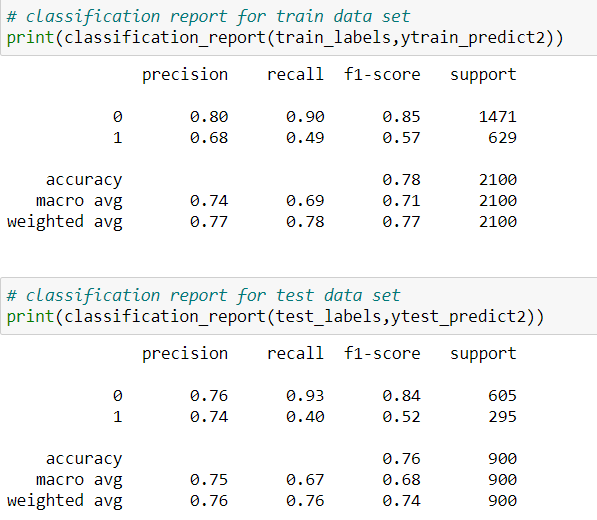
Classification report for CART model:-



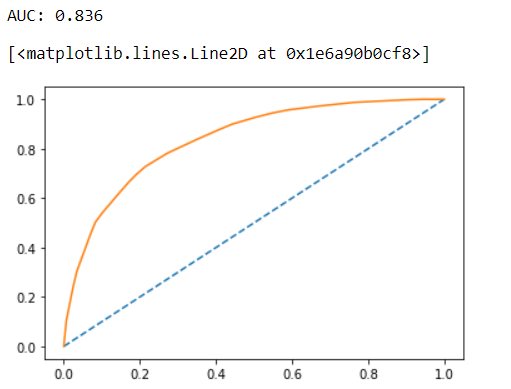
Classification report for random forest model:-



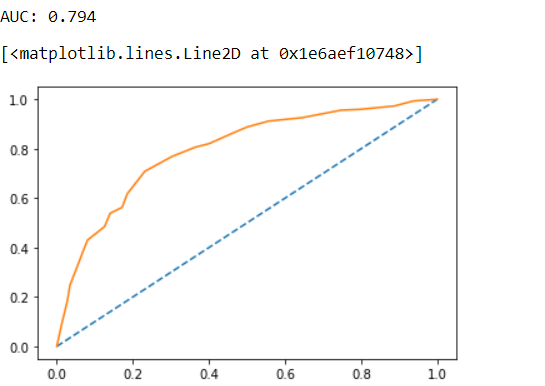
Classification report for MLP model:-



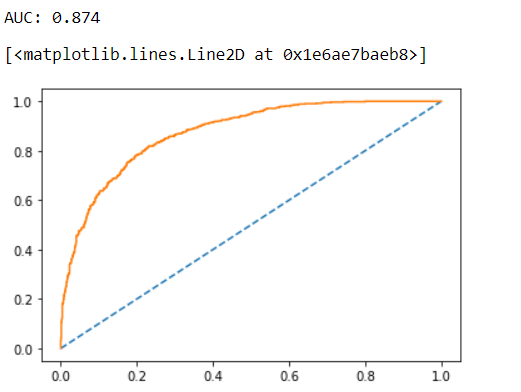
* AUC score and ROC curve for CART model for training data set



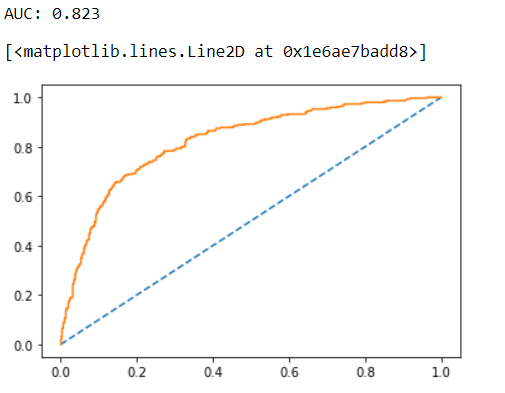
* AUC score and ROC curve for CART model for testing data set



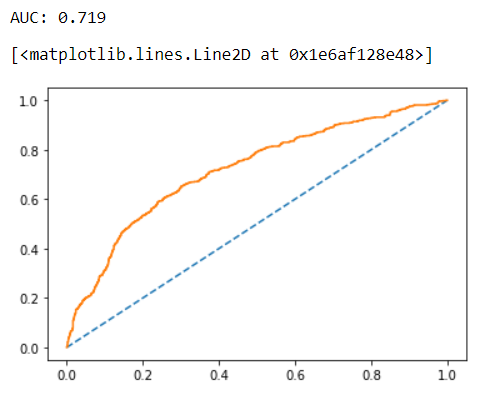
* AUC score and ROC curve for random forest model for training data set



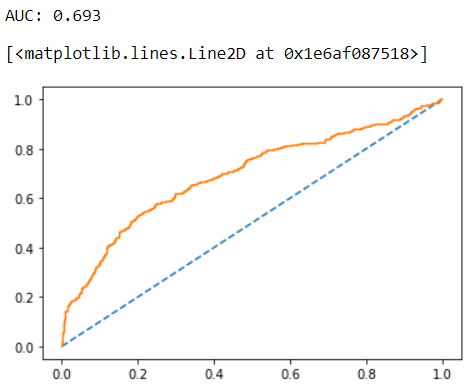
* AUC score and ROC curve for random forest model for testing data set



* ACU score and ROC curve for MLP model for training data set



* AUC score and ROC curve for testing data set



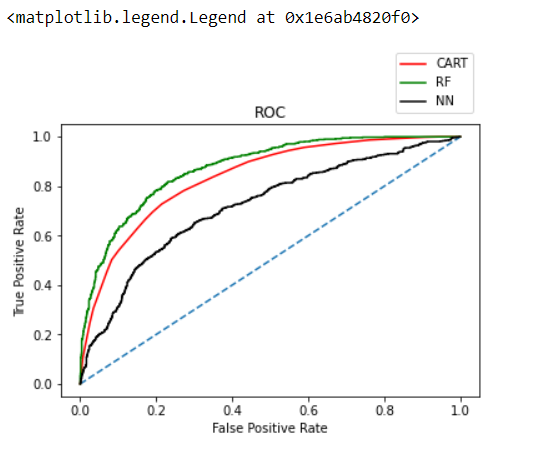
* Observed that while performing random forest the data is over fitted as the accuracy score for training data set is more than testing data set.
* Observed that data is under fitted in MLP model as the training and testing accuracy is same.
* The ideal model for this case study would be decision tree.
* We can get better F1 score and accuracy in random forest and MLP model by trying different combinations of hyper parameters provided for grid search CV and its an iterative process.
* Giving too much value in hyper parameters will increase the system runtime.

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

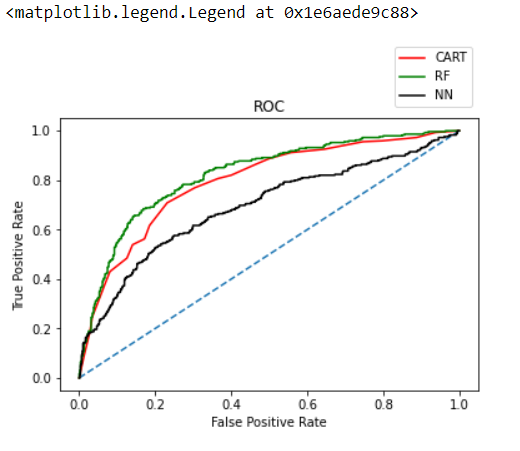
**Solution:-**

Model comparision:-

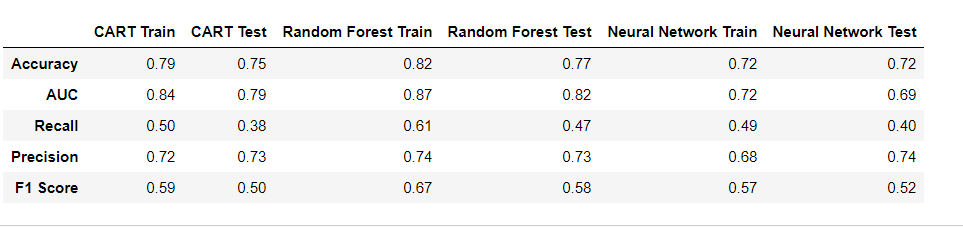
* ROC curve for training data set for all three models prepared:-



* ROC curve for testing data set for all three models prepared:-



* Model comparision based on accuracy, precision, recall and F1 scores for all three models prepared for training and testing data sets



* In my opinion the CART model best suits for this case with decent value of accuracy, precision and recall values.

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

**Solution:-**

Out of the 3 models, CART (decision tree) model has slightly better performance than the random forest and MLP model

Overall all the 3 models are reasonably stable enough to be used for making any future predictions. From Cart and Random Forest Model, the variable change is found to be the most useful feature amongst all other features for prediction. If change is yes, then those customers have more chances of putting claim.

**Recommendations:-**

* The business can charge their clients a higher premium amount who have high probability of claims.
* May tighten their terms and conditions of claiming insurance.
* Business should agree with the touring agency to take care of the clients properly so that they stay well and doesn’t claim much insurance.
* May exclude existing deceases to be covered under insurance.
* May work on the ratio of co-pay and complete payment.
* May acquire more customers as insurance works on the principle of pooling, bigger the pool, its easy for the insurance company to settle the claim smoothly and also with agreed time frame.