DATA MINING PROJECT

SUBMITTED BY

DEEPAK SINGH

**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.

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

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

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

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

1.5 Describe cluster profiles for the clusters defined. 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.

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

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

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

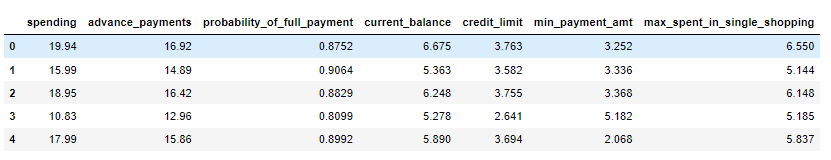
2.4 Final Model: Compare all the models and write an inference which model is best/optimized.

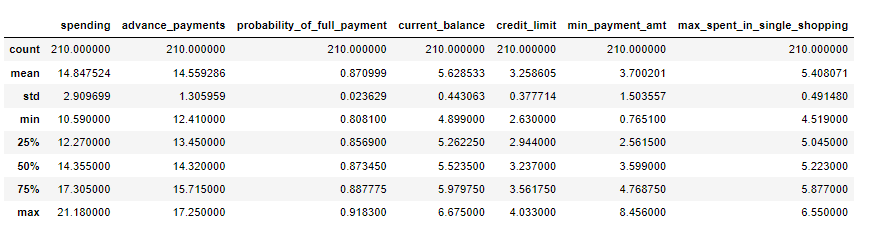
2.5 Inference: Based on the whole Analysis, what are the business insights and recommendations

**Problem 1: Clustering**

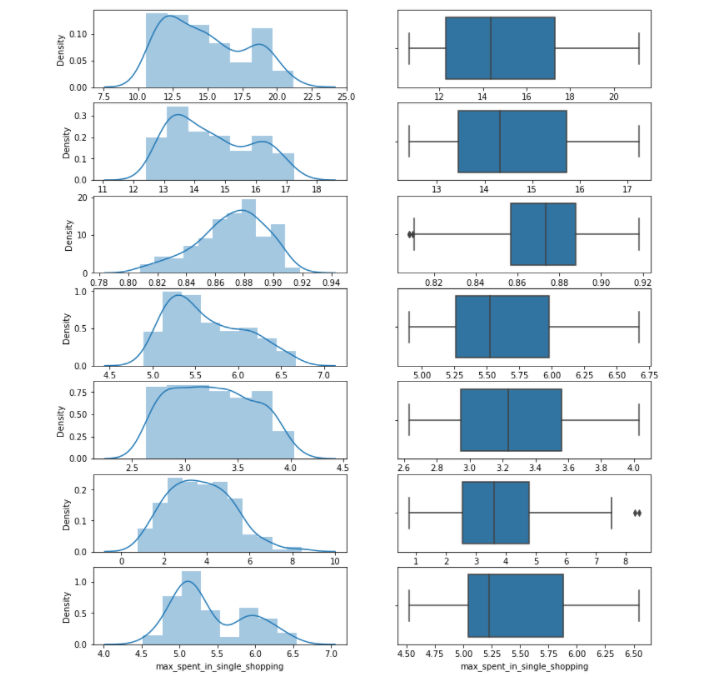
A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

* 1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).
* Top 5 rows of the data set are shown below:



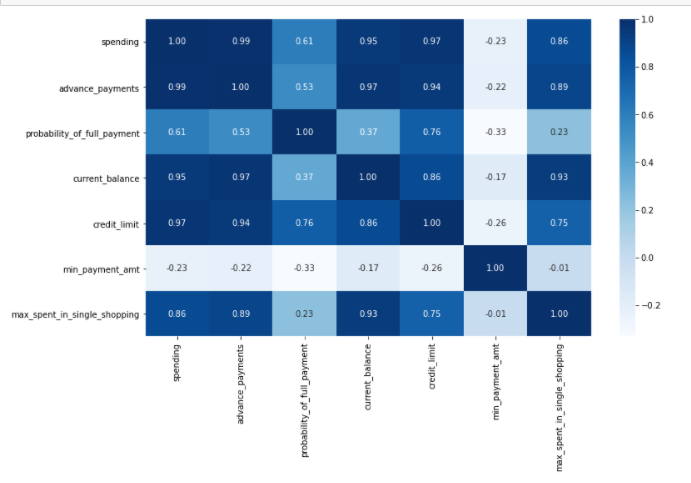
* The given dataset has a shape of (210,7) with no Duplicate & Null Values and datatype as Float64
* Summary of the Dataset:
* Data looks legit as all the statistics seem reasonable. The mean and the median values are almost equal for all the variables. There is no large difference between the 75% and the Max values. By looking at the above to observation we can say that there are no extreme values in the data set

Univariate Analysis: BOX PLOT AND DIST PLOT



No null value found No missing value found Outlier is present only in 2 variables i.e., min\_payment\_amt which means that there are only a few customers whose minimum payment amount falls on the higher side on an average and probability\_of\_full\_payment is in decimal which means there are few customers whose probability to pay full to the bank is on the lower side of average. Since only these variables have a very small outlier value, hence there is no need to treat the outliers. We may conclude that most of the customers have a higher spending capacity, and most of the customers have a higher probability to make full payment to the bank.

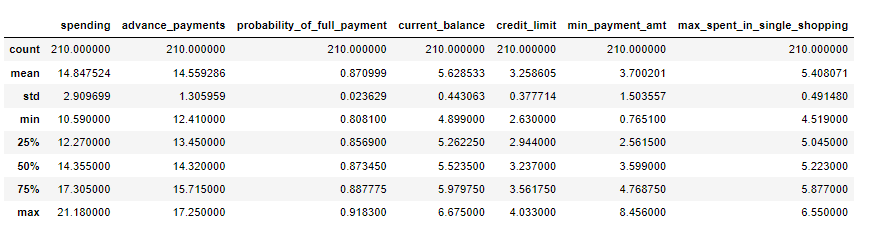
Multivariate Analysis:



From the above Heatmap we can infer that the: spending and advance payments are highly correlated spending and current\_balance are highly correlated spending and credit\_limit are highly correlated Advance\_payment and current\_balance are highly correlated Advance\_payment and credit\_limit are highly correlated Advance\_payment and max\_spent\_in\_single\_shopping are highly correlated current\_balance and max\_spent\_in\_single\_shopping are highly correlated So by this we may infer that the customers who have high credit limit spends more. Also, the customers who pays the money in cash also spends more and the customer who have high Balance amount left in the account to make purchases spends more. min\_payment\_amt is not correlated with any of the variable so it will not affect by any changes in spending, current\_balance or credit limit of the account Probability of full payments is higher for those customers who have a higher credit limit.

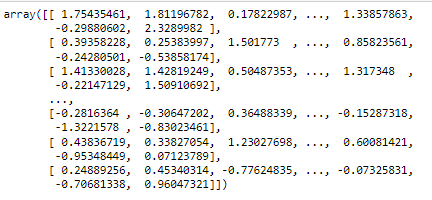
* 1. Do you think scaling is necessary for clustering in this case? Justify

By looking at the summary of data we can see that mean, median, min, max for all variables is distinctly different from each other. Since the hierarchal clustering method uses distance-based computation, scaling is required for unscaled data. Hence it is required to scale and normalize the data. Transformation is required in order to clustering for this case.



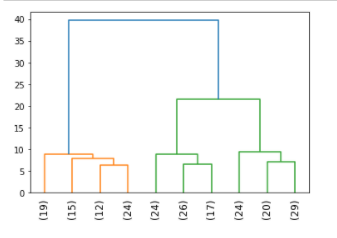
* 1. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Initially we have scaled the data from Standard Scaler.



After performing hierarchical clustering on the scaled data following result is obtained. HCluster method is used to obtain this. Two clusters of “Green and Orange” are obtained through the dendrogram. We find that maximum number of details/customers fall under the red cluster. We have used truncate function with the value P =10 to get the clear output of dendrogram.

**FCluster method is used for Hierarchical Clustering**



In the above output (truncated value) I have drawn a line between 15 and 20 that’s 18, then I see that 3 vertical lines falls under. By using the maxclust method I can see that 3 clusters are good enough.

* 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.

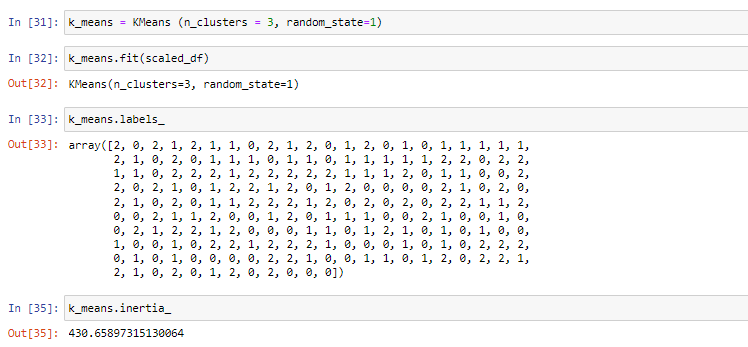
We apply K-means clustering which is a type of unsupervised learning. In this model we need to predetermine the number of clusters and it’s a non-hierarchical model. First, we need to get the optimal WSS plot in order to derive at the optimal number of clusters.

Let’s choose random K value as 3 as show below. We are building the K mean model using K means cluster.

We are fitting the scaled data into K means model. We are able to see that the cluster mapping for the

variables.

* 0 – indicated clusters 1
* 1 – indicates cluster 2
* 2– indicated cluster 3



We can try for different clusters and find the inertia as given below. The larger the drop in WSS it better.

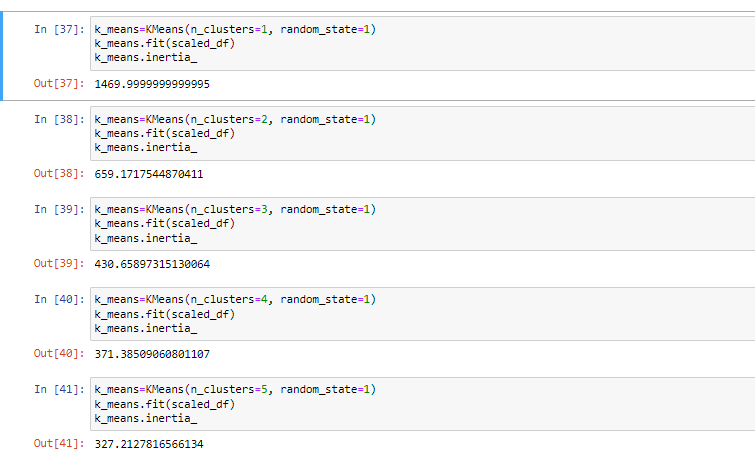
If the drop is not significant the additional cluster is not useful for us.

From 1 cluster to 2 cluster, we have a significant drop close to 900 points

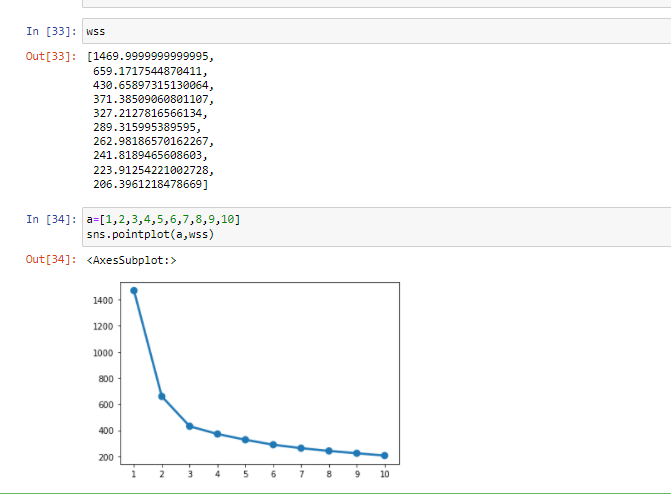
From 2 cluster to 3 cluster, we do have a good drop close to 240 points

From 3 cluster to 4 cluster, it’s not a significant only 50 points drop

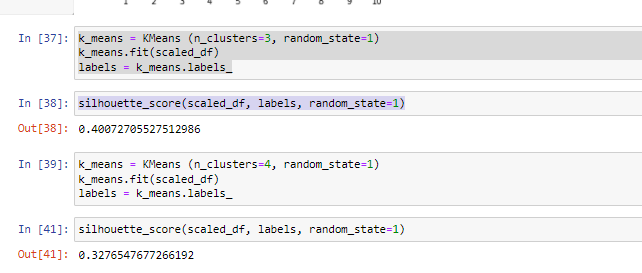
From 4 cluster to 5 cluster, it’s not a significant only 50 points drop



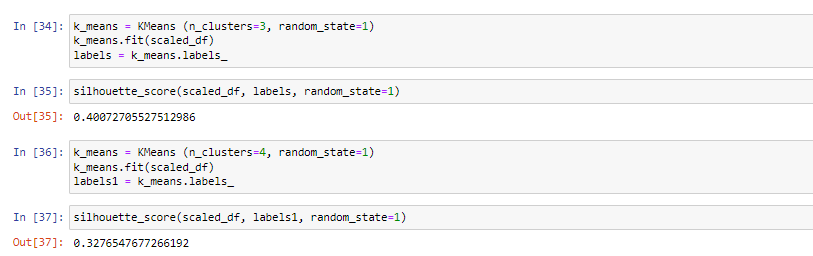
By looking at the drop we can say 3 is optimal for us.



By looking at the graph it’s also evident that drop is significant for 1 ,2 ad 3. post that the drop is very minimal. By using silhouette score and Silhouette analysis we can find that the mapping of each variable to the specific cluster is correct or not.



Silhouette score is better for 3 clusters than for 4 clusters. So, final clusters will be 3



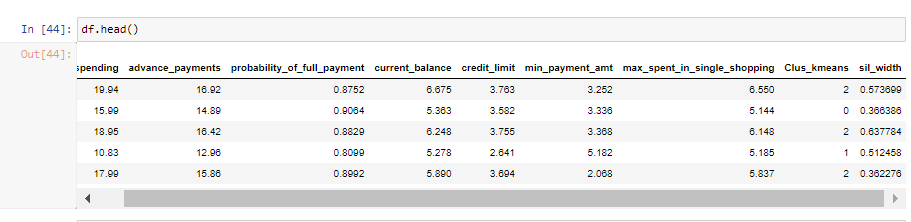
The smallest value of the silhouette width is 0.002 this indicates that no observation is wrongly mapped to a cluster and also all the silhouette width are positive.



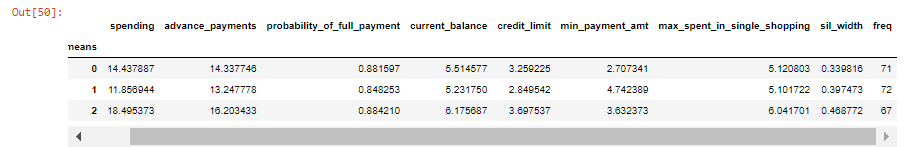
**Hence, we conclude 3 to be the optimal number of clusters.**

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

Two new columns have been added to the data set Clus\_Kmeans and sil\_width. All credit card users are now mapped to one of the three clusters which we have identified (0,1,2 clusters).



By computing the averages of all the customers, we can draw some conclusions.



Looking at above summary we can conclude that:

* Cluster 0: Medium spending group, however the min payment amount is less compared to other least spending group.
* Cluster 1: Least spending group. If you look at the minimum payment amount done by these group people are more than others. Averages of all other parameters are more or less same. We can try increasing the credit limit of the customers who fall under the cluster 1. Since the average amt spent in a single shopping is more are less equal to cluster 0.
* Cluster 2: Premium group who spends more money. Credit limit and all other parameters are also relatively high compared to members of other clusters. The minimum payment amount is less compared to the least spending group. Bank can look into that to increase the minimum

**PROMOTIONAL STRATEGIES:**

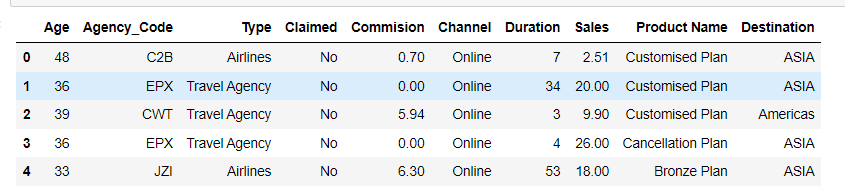
Those who are clustered under 2(3rd cluster) are seen as elite customers who spend more monthly and all other parameters are also relatively high compared to members of other clusters. It is seen that minimum paid by the customer while making payments for purchases made monthly (in 100s) is high for those under the second (1) cluster. The credit limit for those under (1) can be increased as a promotional strategy as their averages of all parameters are relatively same compared to others. Max amount spent on a single purchase is more or less the same for those under 0 and 1 clusters (first and second). Customers can spend on different products. Customers under third cluster (2) seem to spend more, but their minimum payment is less than those of second cluster customers. Bank shall increase their slab of min payment amount. Advance payments made by second cluster customers are less, which the bank shall look into payment amount for the premium group.

Problem 2: CART-RF-ANN

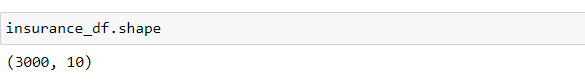
An Insurance firm providing tour insurance is facing higher claim frequency. The management decides to collect data from the past few years. You are assigned the task to make a model which predicts the claim status and provide recommendations to management. Use CART, RF & ANN and compare the models'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

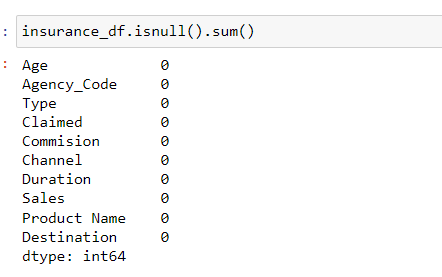
Data Head:



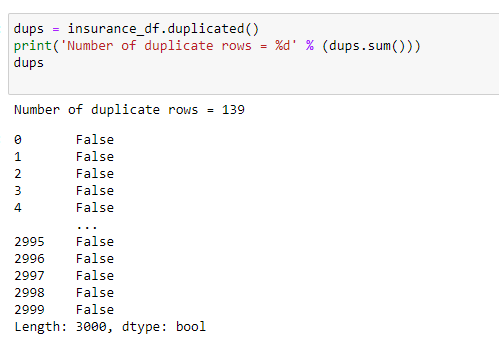
Data Shape:



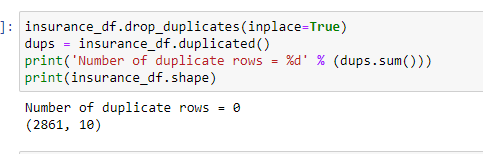
Check for null values in columns:



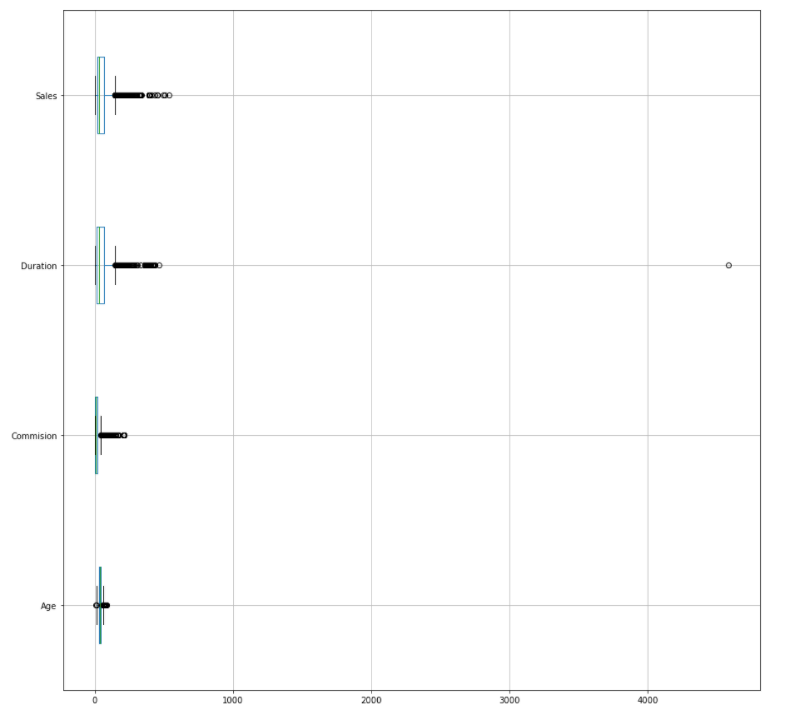
Check for duplicate data:



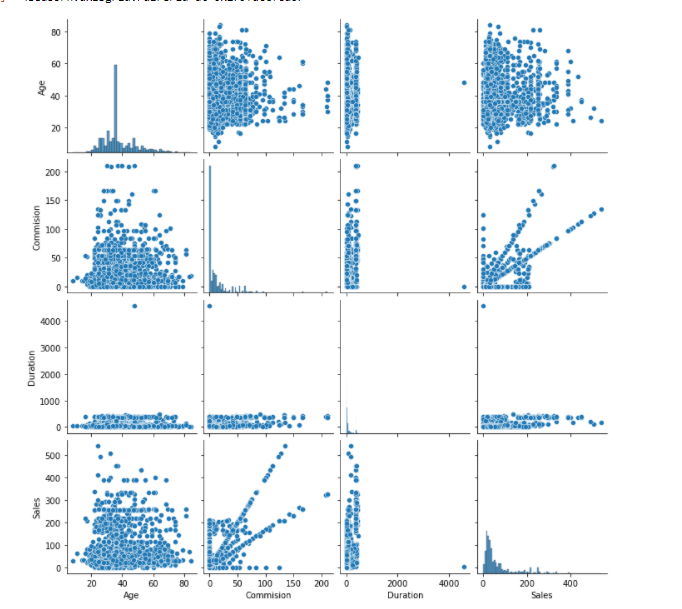
Removing Duplicates:



Checking for outliers:

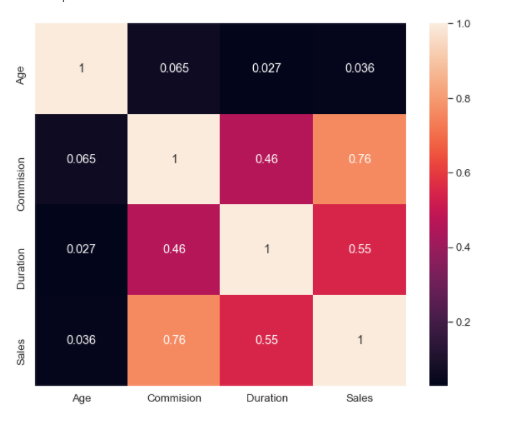


Checking pairwise distribution of the continuous variables:



Do the count plot (pls refer Jupyter Notebook)

Checking for Correlations:

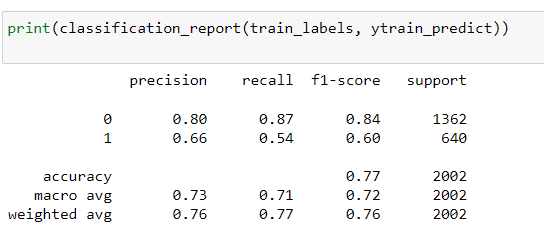


2.2 Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

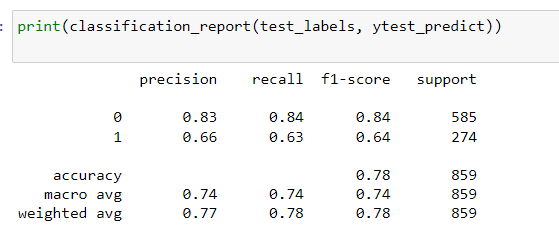
Please refer the Jupyter Notebook for splitting of data for CART, RF & ANN. Post that we have derived the classification Model:

Classification Model for Cart:

Train Data:

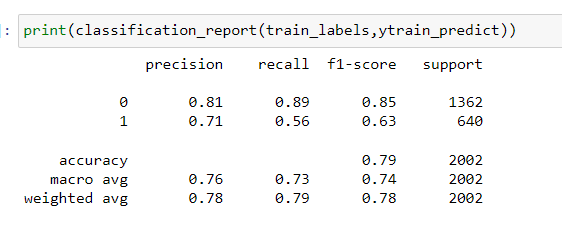


Test Data:

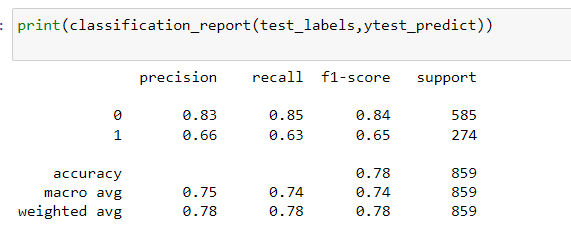


For RF:

Train Data:

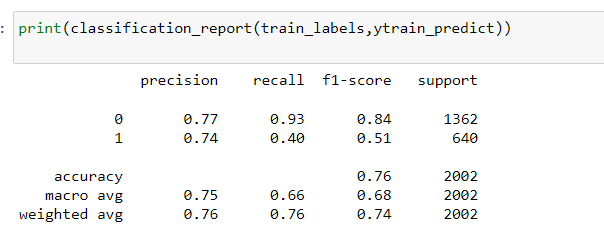


For Test Data:

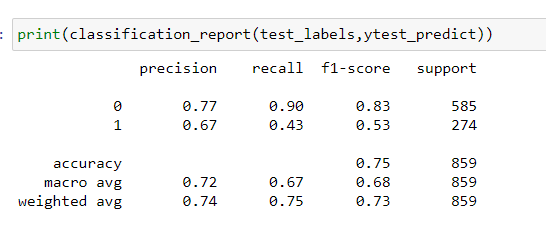


For ANN:

Train Data:



For Test Data:

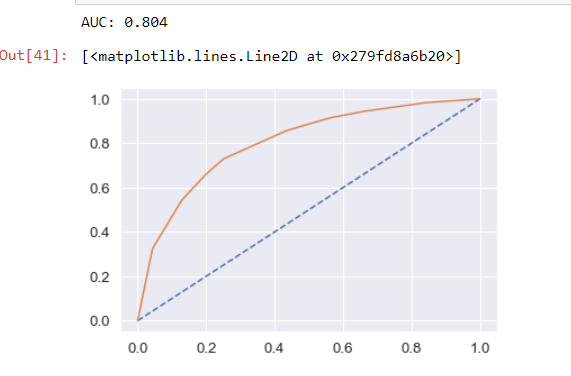


2.3 Performance Metrics: Comment and Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score, classification reports for each model.

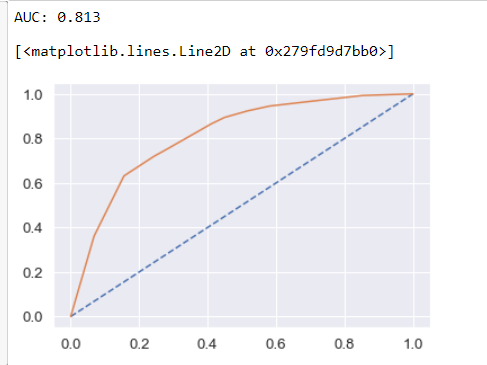
CART:

AUC/ROC for Training data is given below, ROC is a curve represents TP rate and FP rate

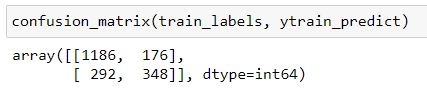
Train Data:



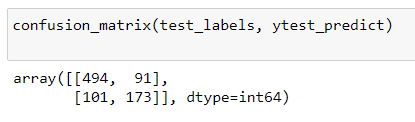
Test Data:



Confusion matrix for the training data is given below:



Confusion matrix for the test data is given below:



Cart Conclusion: By looking at the confusion matrix, AUC/ROC curve and classification report we can derive the below:

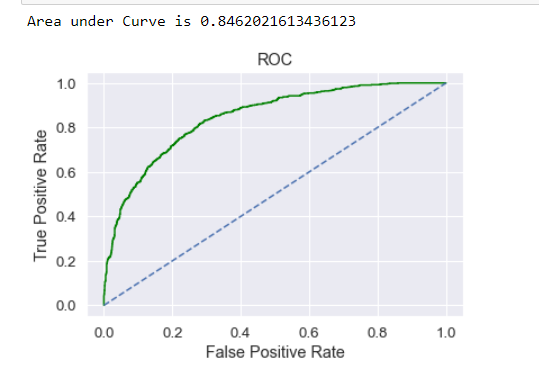
Train Data: AUC:80% Accuracy: 76% Precision: 66% f1-Score: 60% Recall: 54%

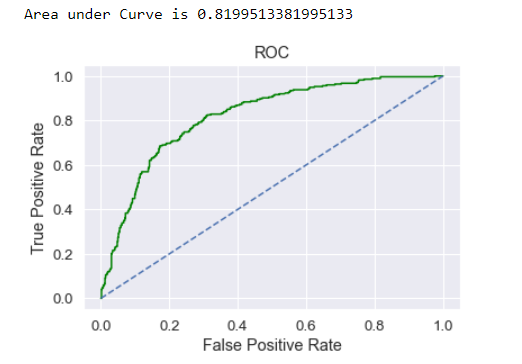
Test Data: AUC: 81% Accuracy: 78% Precision: 66% f1-Score: 64% Recall: 63%

Training and Test set results are almost similar, and with the overall measures being moderate, the model is a good model.

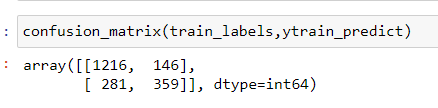
RF

AUC/ROC for Training & Test data is given below, ROC is a curve represents TP rate and FP rate

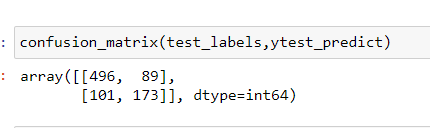




Confusion Matrix Train:



Confusion Matrix Test:



RF Conclusion: By looking at the confusion matrix, AUC/ROC curve and classification report we can derive the below:

Train Data: AUC: 84% Accuracy: 79% Precision: 71% f1-Score: 63%

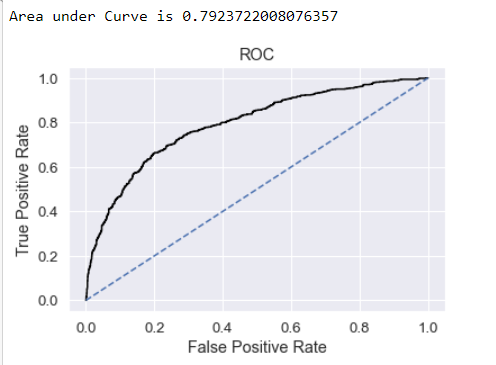
Test Data: AUC: 81% Accuracy: 78% Precision: 66% f1-Score: 65%

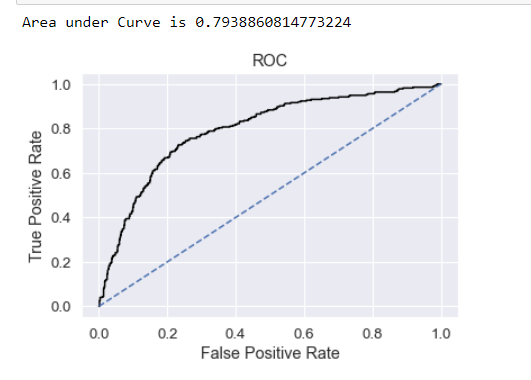
Training and Test set results are almost similar, and with the overall measures high, the model is a good

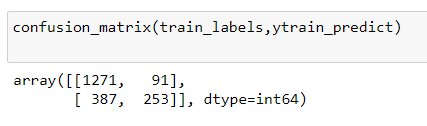
modeAgency\_code is again the most important variable for predicting customer insurance claim

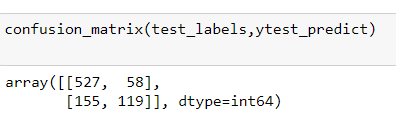
ANN

AUC/ROC for Training & Test data is given below, ROC is a curve represents TP rate and FP rate:









ANN Conclusion: By looking at the confusion matrix, AUC/ROC curve and classification report we can derive the below:

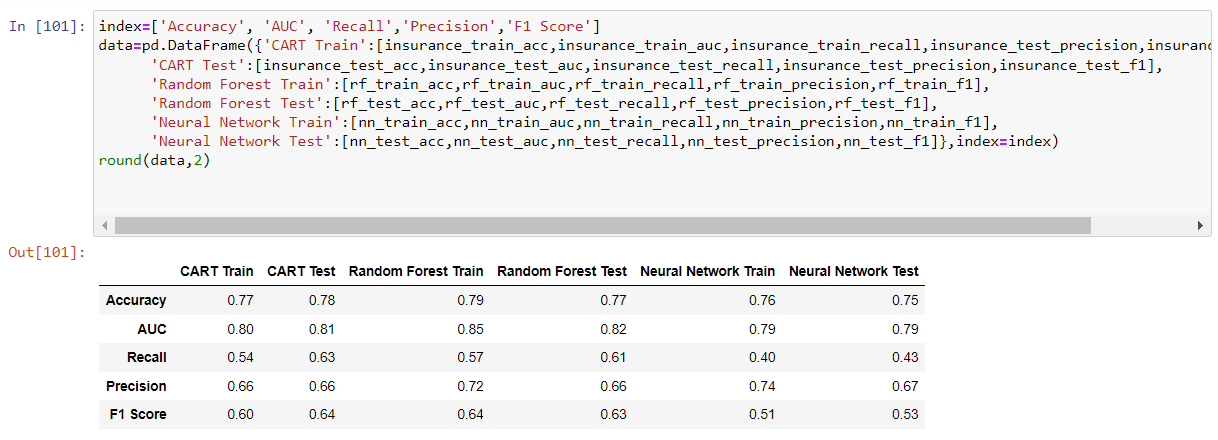
Train Data: AUC: 76% Accuracy: 79% Precision: 74% f1-Score: 51%

Test Data: AUC: 75% Accuracy: 78% Precision: 67% f1-Score: 53%

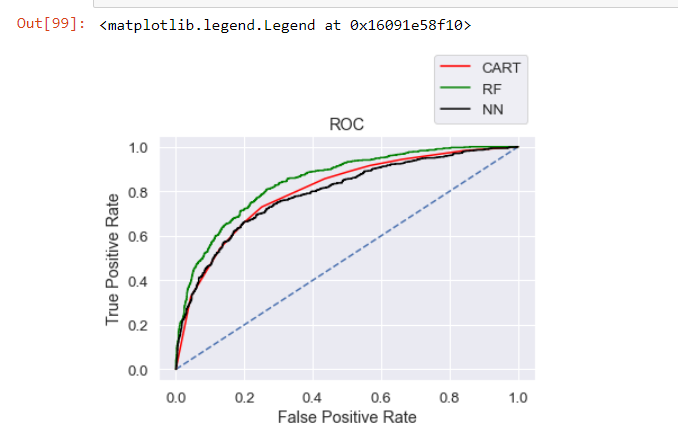
Training and Test set results are almost similar, and with the overall measures high, the model is a good model.

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

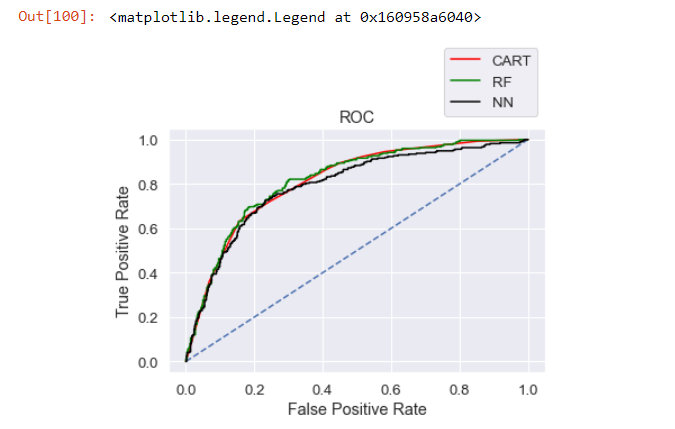
Comparison performance metrics from all 3 models –



ROC curve for all 3 models on Training Data –



ROC curve for all 3 models on Training Data –



We are able to see that Recall and FI percentage are good for CART, RF compared to Neural networks and also its above 0.5. For ANN the value is less than 0.5, which is not good. ANN is not a good model for this data set. When comparing CART and RF by using single decision tree we are able to get better result than generating multiple trees. Since we are getting good results with minimal tree, we tend to go for CART for the given data set and avoid Random Forest as well.

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

CART model has relatively high Recall, FI core compared to another model. Precision and AUC is more Random Forest compared to another model. Recall is the popular measure over all, accuracy will be checked when the target column is balanced. Hence CART seems to be a better model as a conclusion. Channel and Type seems be the major factor which business needs to concentrate. These two factors play a primary role in predicting the future claims.

