**Problem 1: Clustering**

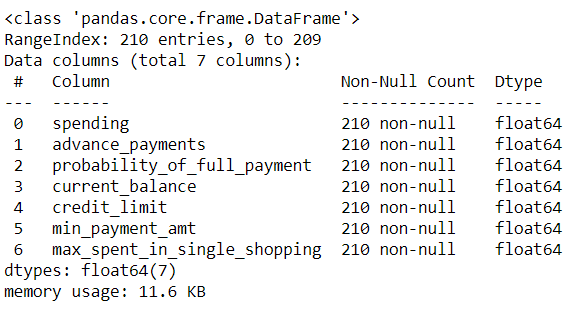
Problem summary: To do customer segmentation of a leading bank based on their transaction behavior in their credit cards.

Given variables in the data: spending, advance\_payments, probability\_of\_full\_payment, current\_balance, credit\_limit, min\_payment\_amt, max\_spent\_in\_single\_shopping.

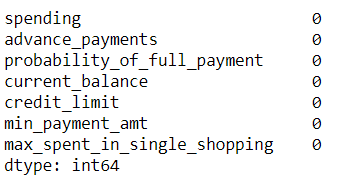
**1.1** Read the data and do exploratory data analysis. Describe the data briefly.

Answer:

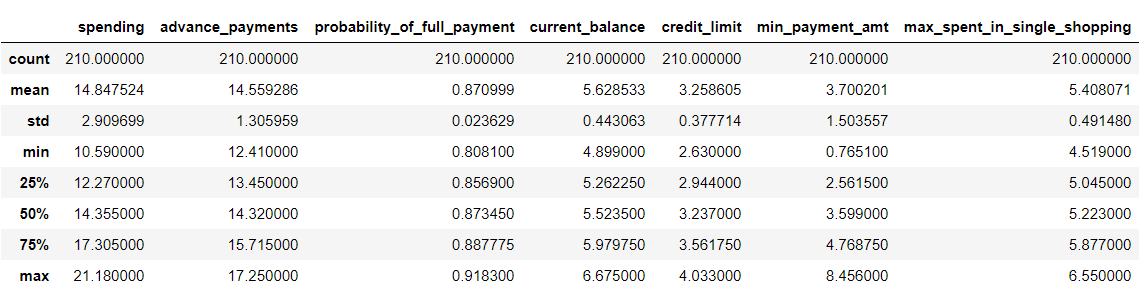
* Data has 210 entries of 7 columns with all the data having a decimal value:



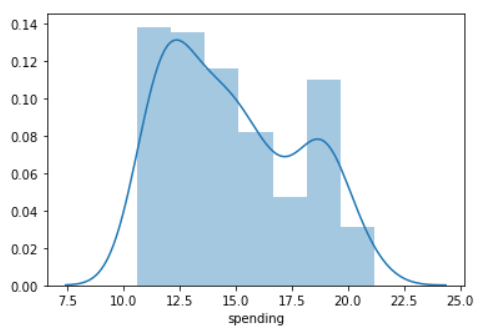
* We checked for any inconsistencies in the data using the value count function and we did not find any.
* We then checked for any missing data and there were none:



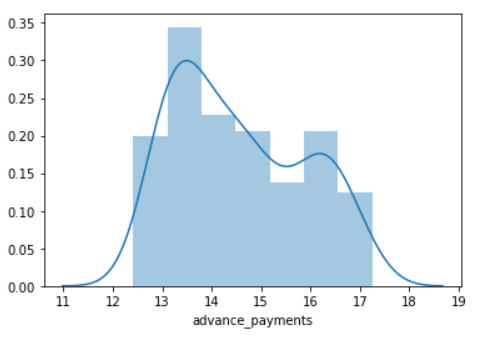
* Using the describe function we were able to get the following information on the distribution of the data in each of the columns:



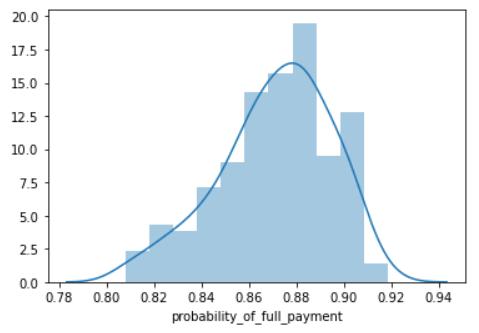
* **Univariate Analysis:**



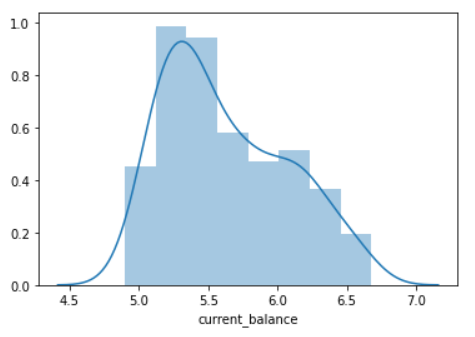
* The spending variable is slightly right skewed.
* A spike in the data is observed between the ranges of 18.5 to 19.5 which means 18,500 and 19,500.



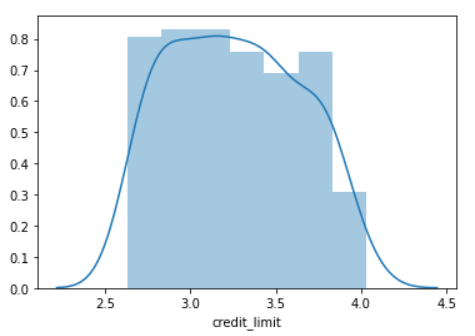
* The advance\_payments section of the data is also slightly right skewed with a spike being observed between 16 – 17.



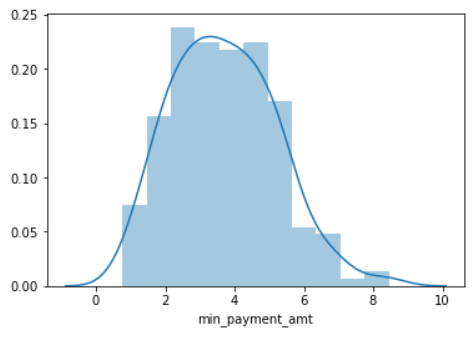
* Probability of payment done in full by the customer to the bank is a left skewed distribution which is a good thing that means that there will be no dues or carry over balance present.



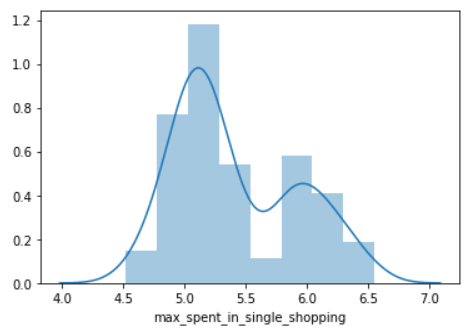
* The variable ‘current\_balance’, is slight right skewed. This is a variable that would be better if the skewness is to the left.



* The Limit of the amount in credit card (10000s) has a distribution that is symmetric.

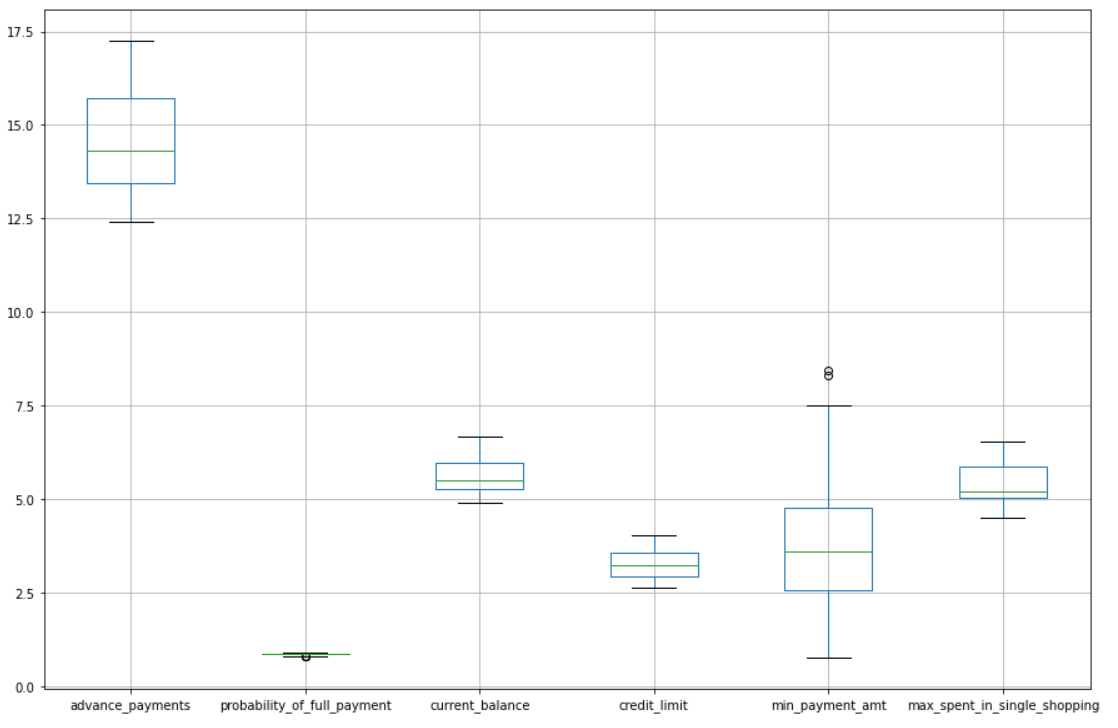


* minimum paid by the customer while making payments for purchases made monthly (in 100s) is also a symmetrically distributed data.



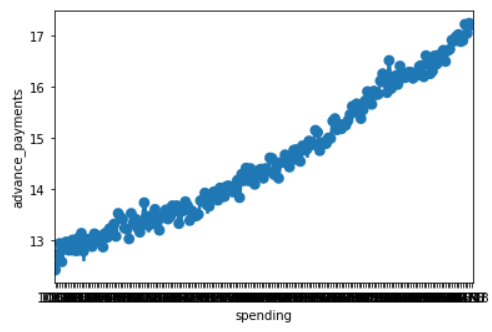
* Slight right skewness can be observed in this variable of Maximum amount spent in one purchase (in 1000s) with a spike in the 5.9 – 6 mark.

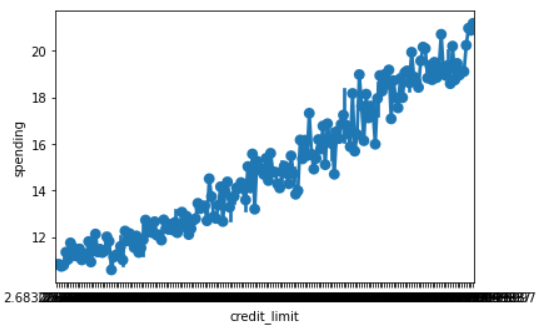
**Boxplot of the variables:**

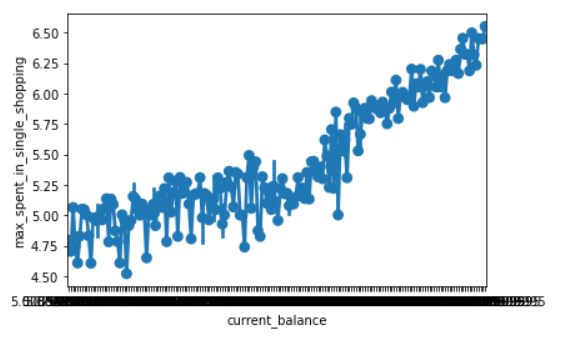


* None of the variables have outliers except for minimum paid by the customer while making payments for purchases made monthly (in 100s). The outliers are negligible as they will not cause any significant variation in our analysis.
* The probability of full payment can be ignored as it is a probability between 0 and 1.
* There were also no duplicate entries found in the dataset.

**Bivariate Analysis:**





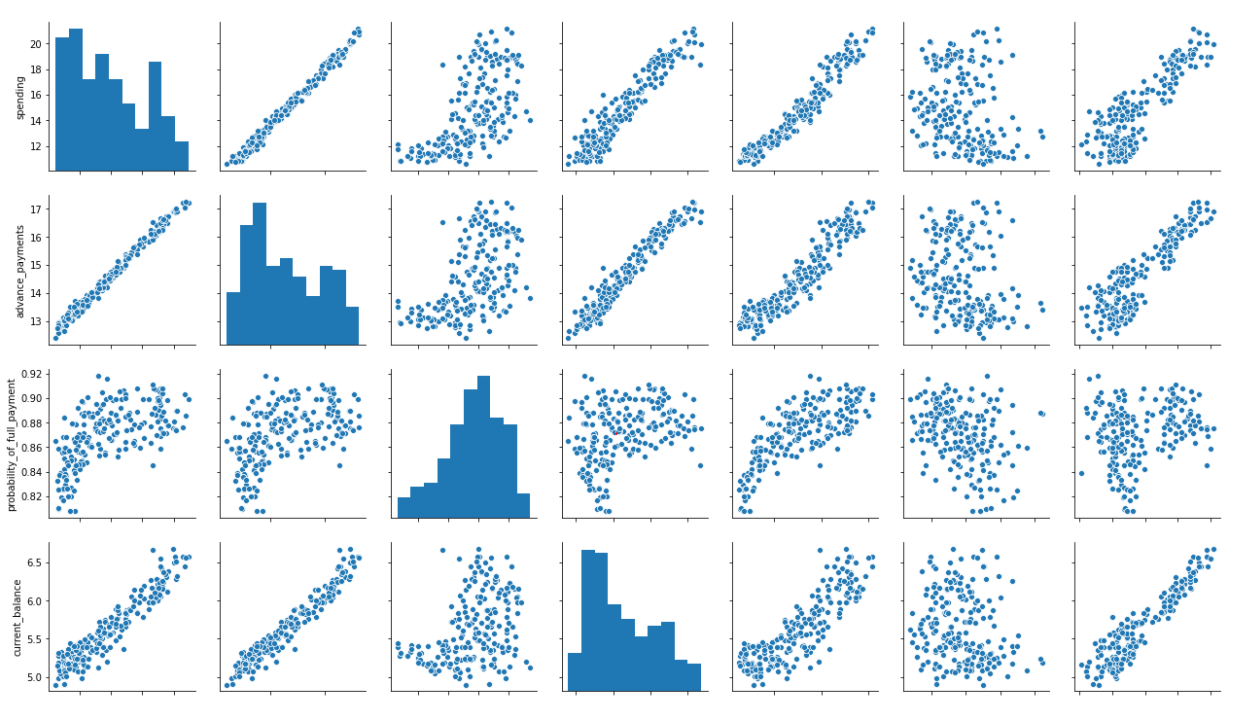


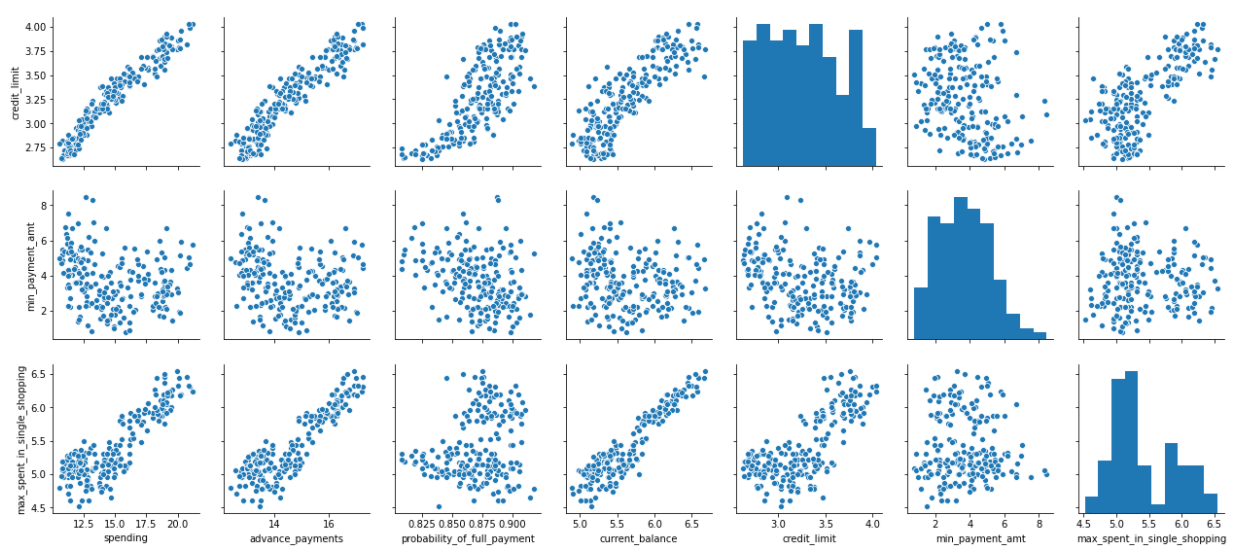


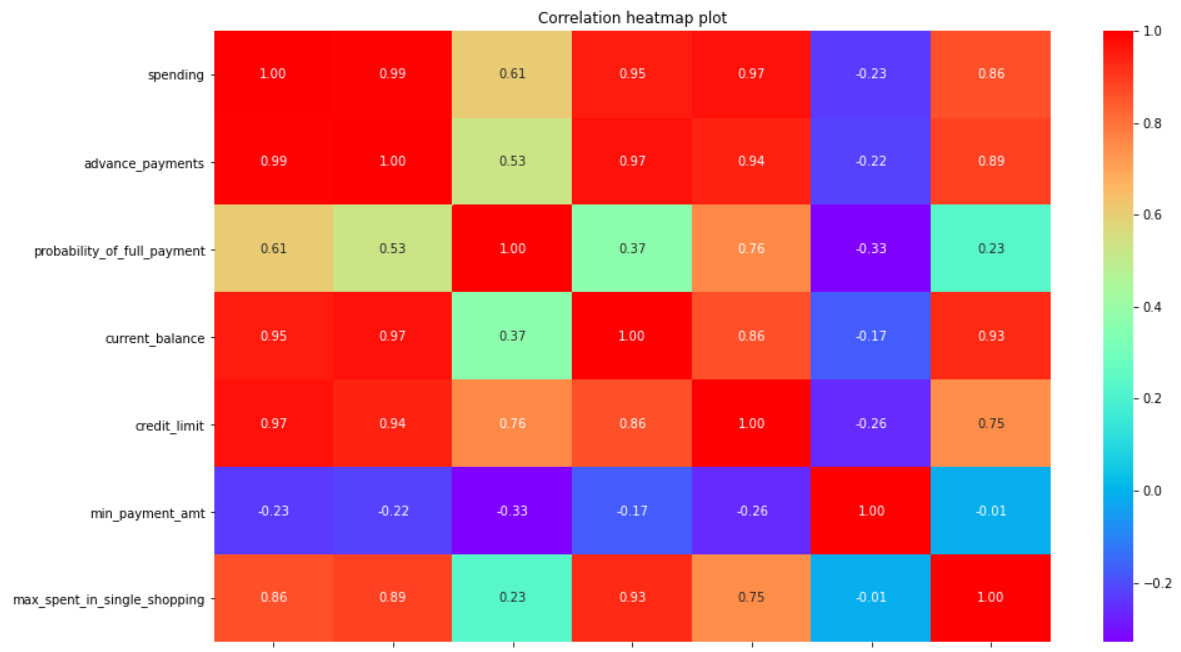
* The variables spending, advance payments and credit limit have a perfectly linear correlation as seen.
* The variables max spent in a single shopping, credit limit and current balance are also linearly correlated with some abnormality.

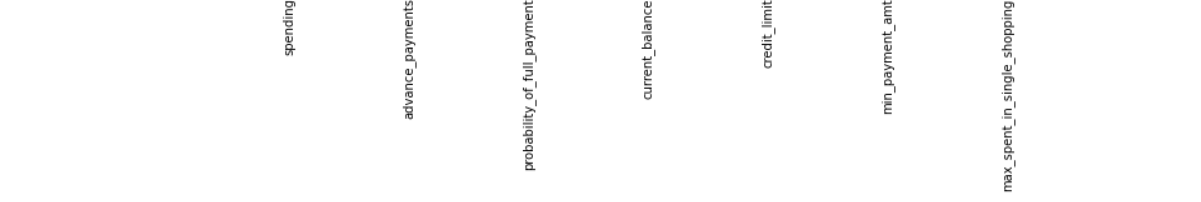
**Multivariate Analysis:**

**Correlation Pairplot:**





**Correlation Heatmap:****



* Most of the variables are linearly correlated to each other.
* The correlation between each of the variables in the dataset is high.
* Extremely high correlation exists between the following variables: spending, current\_balance, advance\_payments, credit\_limit and max\_spent\_in\_single\_shopping.
* The variables: prob\_of \_full\_payment and min\_payment\_amount are the only two variables which have a low correlation with the other variables. We will see the results with and without these variables in the further sections of our analysis.

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

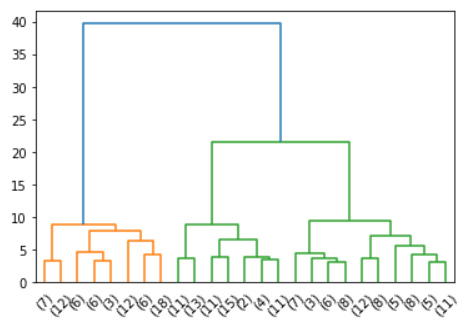
Scaling is essential on this data due to multiple reasons:

* Some of the variables in the dataset have an average two digit figure whereas some have a single digit. One variable has only values between 0 and 1 as it represents a probability.
* Since we will perform algorithms like K-nearest neighbors and PCA to find out patterns and insights, scaling becomes an essentiality.
* Even a slight variation of results while using the clustering algorithms presents a large effect on the outcome because the variables are representing different magnitudes of data.

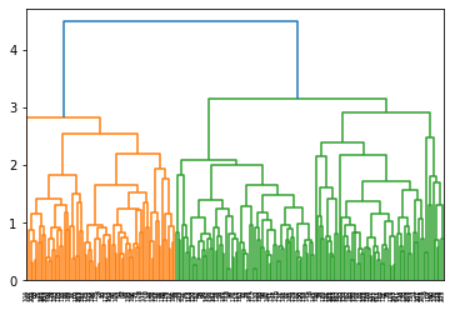
This is seen in the data dictionary of the problem statement where we have the columns representing different magnitudes. We have data distributed between 100s, 1000s and 10000s.

**Questions 1.3, 1.4 and 1.5 will be addressed in the following sections of our analysis:**

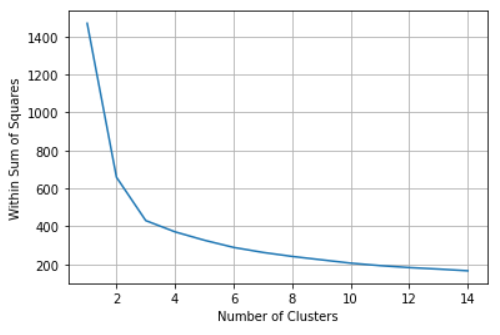
* We first scale the data using the z score scaler to perform hierarchical clustering.
* The dendrogram using the wards linkage method of the last 25 merges looks like this:



* Then the clusters were extracted using the max clust method.
* A dendrogram was created using the average linkage method and was rejected as the clusters were not well defined as in the case of Ward’s method.

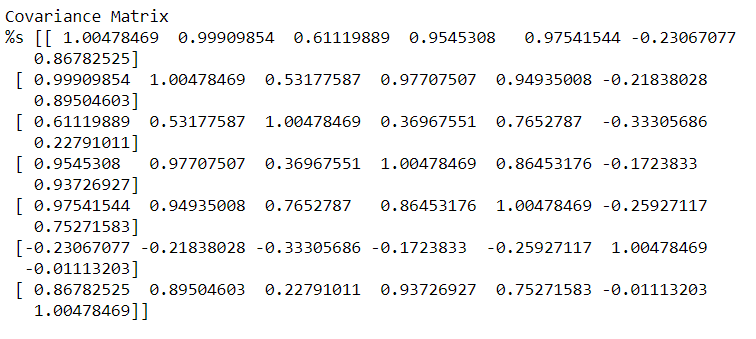


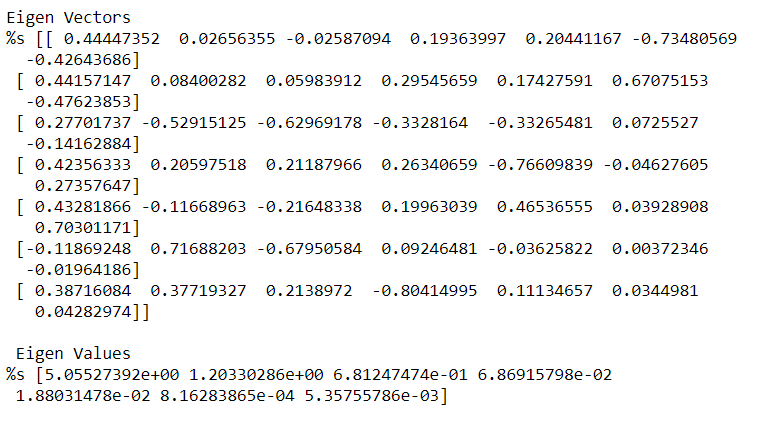
* K-Means Clustering was performed, the labels were created, the inertia value for 3 clusters was computed to be 430.658 which was greater than the inertia value for 4 clusters (371.385) using the same data.
* The Elbow plot for the same:

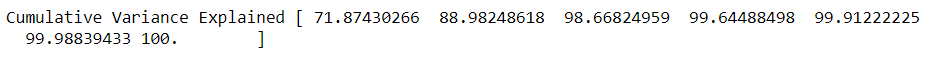


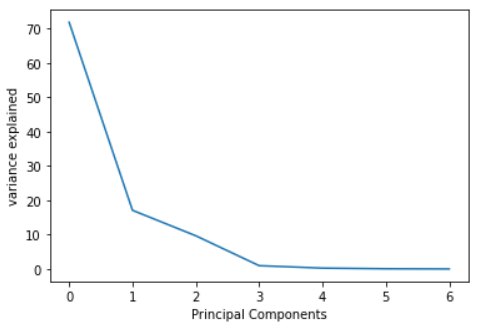
* The Silhouette score for 3 clusters is 0.40072 and the Silhouette width is 0.0027.
* The Silhouette score for 4 clusters is 0.32765 and the Silhouette width is -0.0538.
* Agglomerative clustering was done on the data using the Euclidean distance and the average linkage method. The clusters using the Hierarchical clustering, K-Means clustering and the agglomerative clustering techniques were appended to the data.

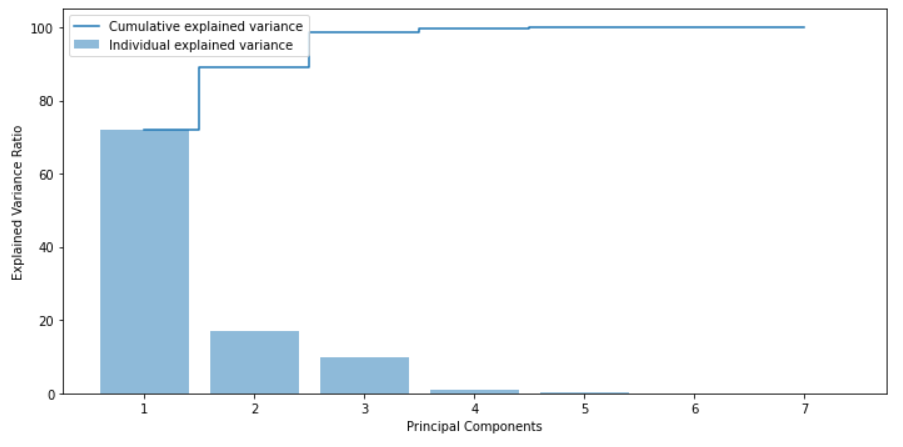
**Principal components:**



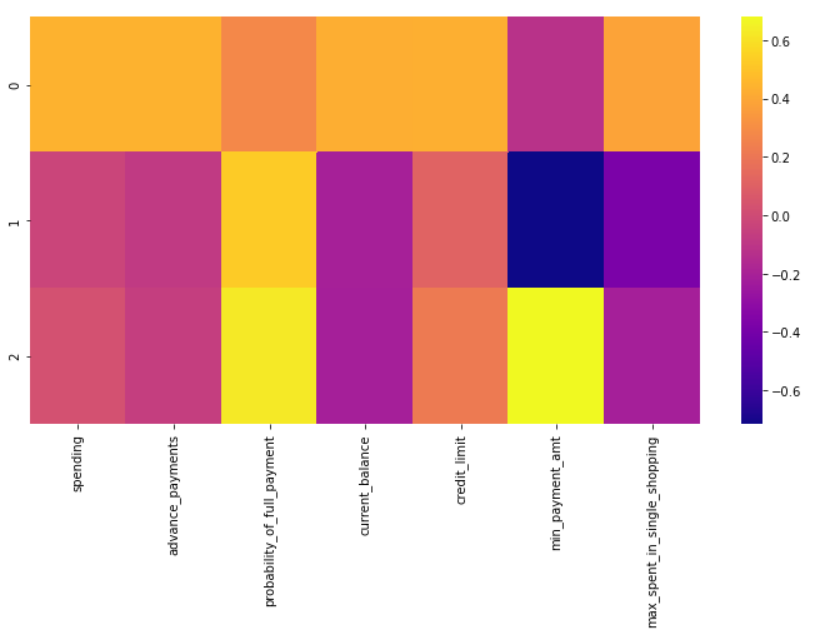






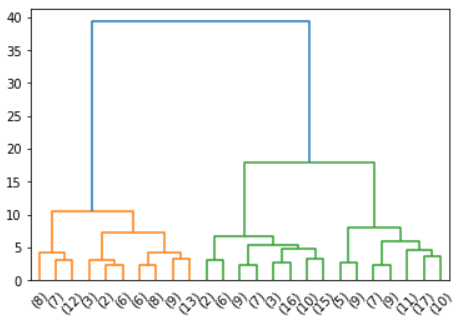




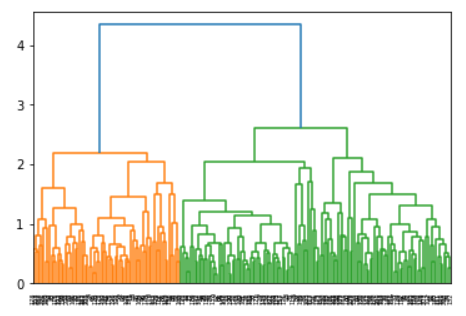


**Clustering without the column: Probability of Full Payments:**

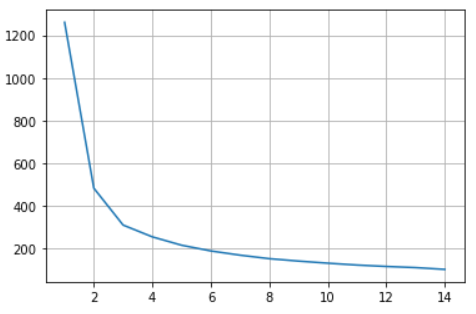
* After scaling, we perform hierarchical clustering using the Ward’s method to create a dendrogram:



* The distance of the two clusters in the green color has slightly increased as compared to the previous dendrogram with the inclusion of the variable that has been excluded in our analysis here.
* We created a dendrogram using the average method as well but the separation of the clusters has not been acceptable:

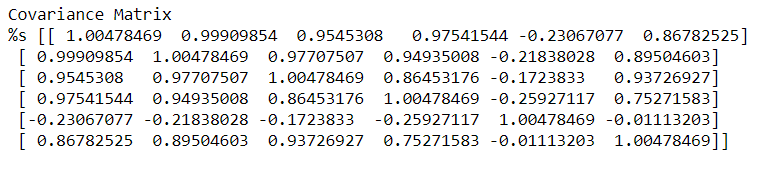


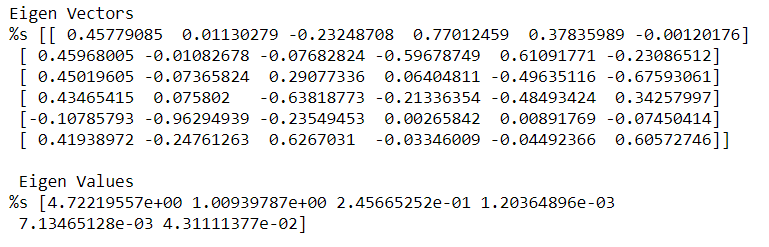
* K-Means was performed using 3 clusters. The Elbow plot for the same was computed:



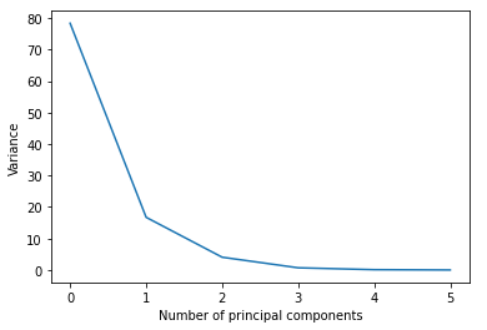
* It can be seen from the above plot that the curve has become steeper than before. It can be inferred from this that the variable probability of full payment was more of a noise in our analysis.
* The Silhouette score for 3 clusters was found to be 0.43598 which is more than the Silhouette score for 3 clusters with the column 'Probability of Full Payments' which was 0.40072.

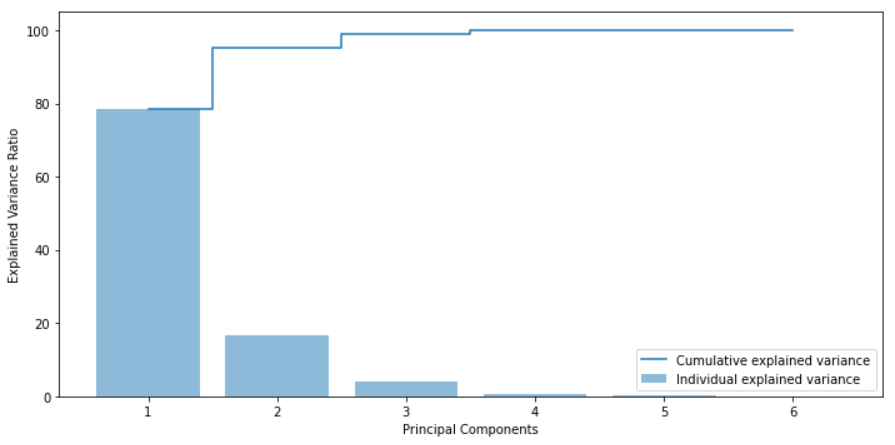
**Principal Components for the data without the column, 'Probability of Full Payments':**



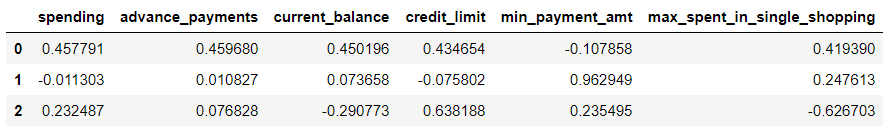






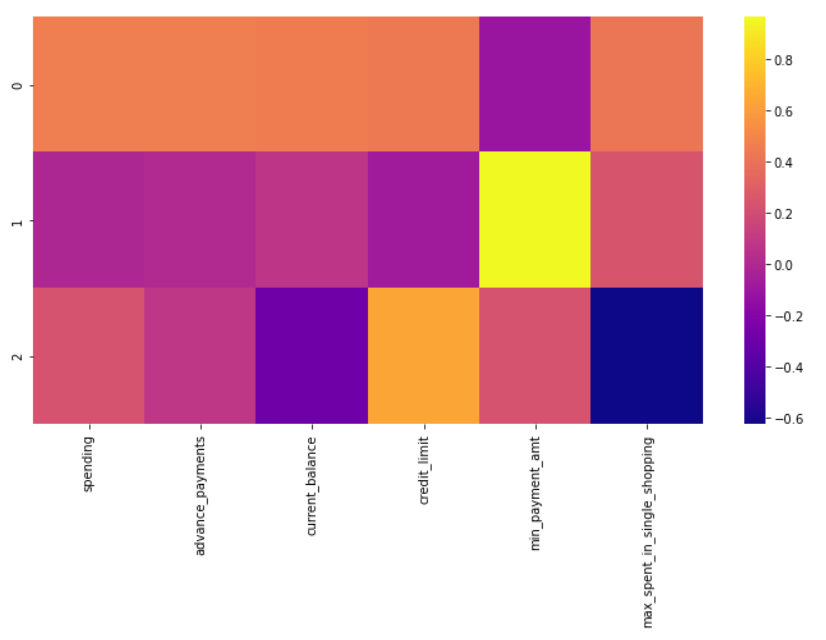


* PCA components for n = 3 were extracted and a dataframe was formed:



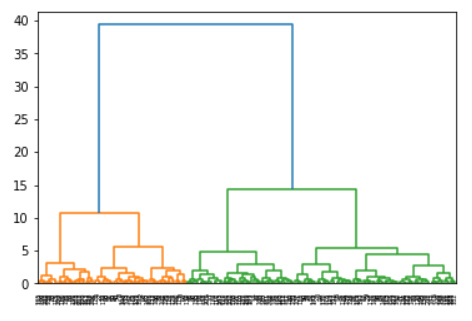
* A heatmap for the principal components and their respective correlation values was shown:

It is interesting to note that only the minimum payment amount is not in line with the behavior of the other variables in each of the principal components. It is showing an opposite correlation.

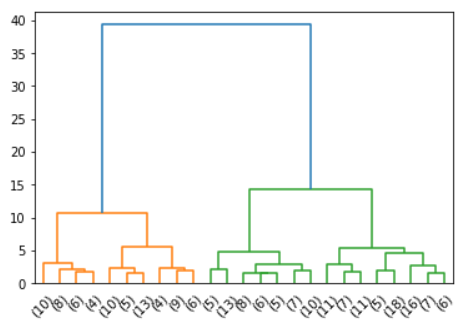


**Clustering without the column: ‘Probability of Full Payments’ & 'min\_payment\_amt':**

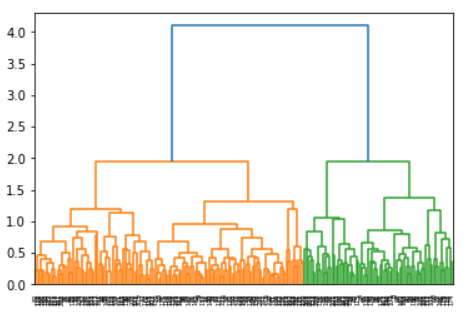
* After scaling, we perform hierarchical clustering using the Ward’s method to create a dendrogram:



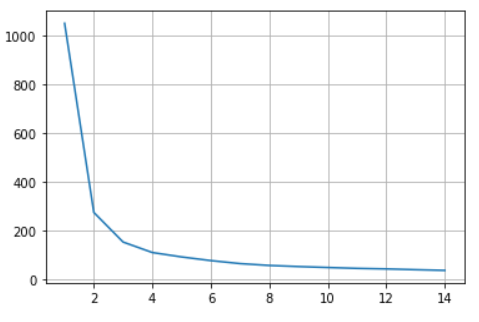
* The dendrogram shows us 4 clusters. The depth of the 4th cluster is not much as the quantity of observations provided in our dataset is minimal.



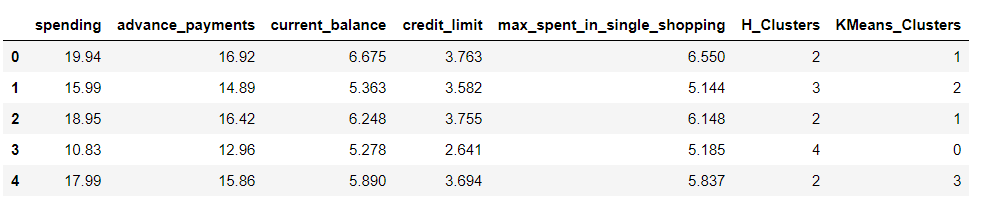
* We also performed hierarchical clustering using the average method and the following results were obtained:



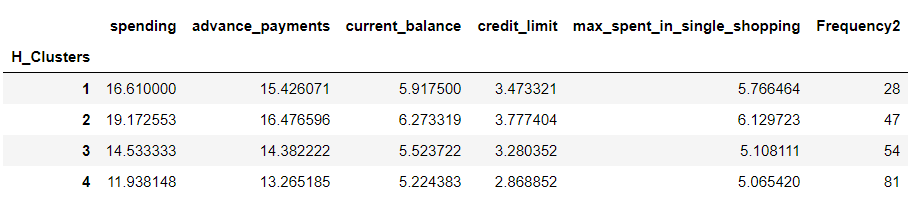
* K-Means Clustering was performed for 4 clusters for which the inertia and labels were computed.
* The Elbow plot for the same was shown and the silhouette score for the same was computed.



* The clusters from the hierarchical clustering (ward’s method) and K-Means clustering were appended to the data to perform cluster profiling:



**Cluster Profiling (using only the clusters from the hierarchical clustering technique):**



* The following shows the number of customers in each of the clusters 1,2,3 & 4:

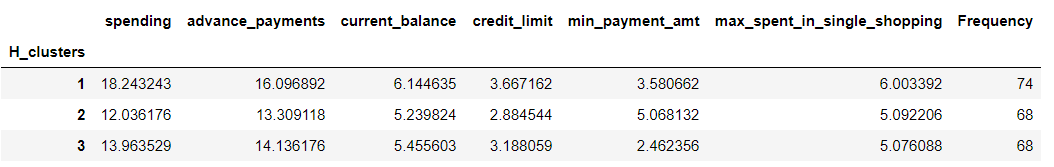


* Cluster 2 indicates the financially more stable customers as the amount spent by the customer per month, amount paid by the customer in advance by cash, the balance amount left in the account to make purchases, credit limit and the maximum amount spent in one purchase is high.
* A credit card having a requirement of a high monthly expenditure and in return reward points with attractive redemption options for those reward points could be presented to the group of customers belonging to cluster.
* Cluster 4 indicates financially low customers with the least values in the parameters mentioned above. It is however interesting to note that these are the customers who have highest numbers in the section: minimum paid by the customer while making payments for purchases made monthly.
* The credit limit for these customers is the lowest.
* Products with low price options could be presented to this group of customers belonging to cluster 4 as their current balance is the least in the data. For example: with a low minimum deposit saving scheme.
* Clusters 1 and 3 indicated the group of customers who can be classified in between the clusters 2 and 4.
* They could be attracted into buying a loan from the bank with low interest rates. This could be aligned with longer repayment periods to avoid burden on the customers belonging to these clusters.

**Cluster Profiling for 3 clusters:(removing : probability\_of\_full\_payment)**

* Here are the number of customers in each of the 3 clusters:





* The minimum payment amount cannot be underestimated as the value is high for the customers who belong to cluster 2. It is interesting to note that these are the customers who have low values in spending, advance payments, current balance and credit limit.
* The probability of full payment can be removed from the data as the customers have an average of 0.8 as the probability of full payment irrespective of the cluster they belong to. This has been observed in the analysis using the variable and is available in the python computation sheet for reference. This variable has shown to be more of a noise in the data.
* It is however pleasant to see the average probability of full payment for all the customers is 0.8. This indicates the low number of customers who are prone to default in their payments to the bank.

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

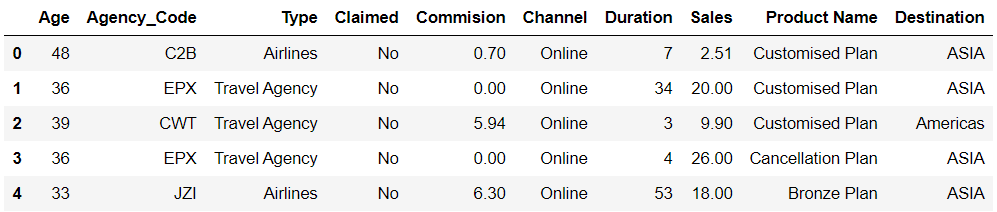
**Problem Summary:** To predict the insurance claim status using models like CART, Random Forest and ANN thereby provide recommendations to the management.

Independent variables in the data: Agency\_Code, Type, Channel, Product, Duration, Destination, Sales, Commission.

Dependent variables in the data: Claimed

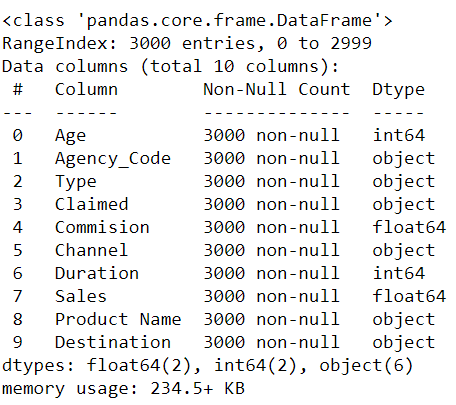
**2.1** Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it.

The head of the given data looks like this:

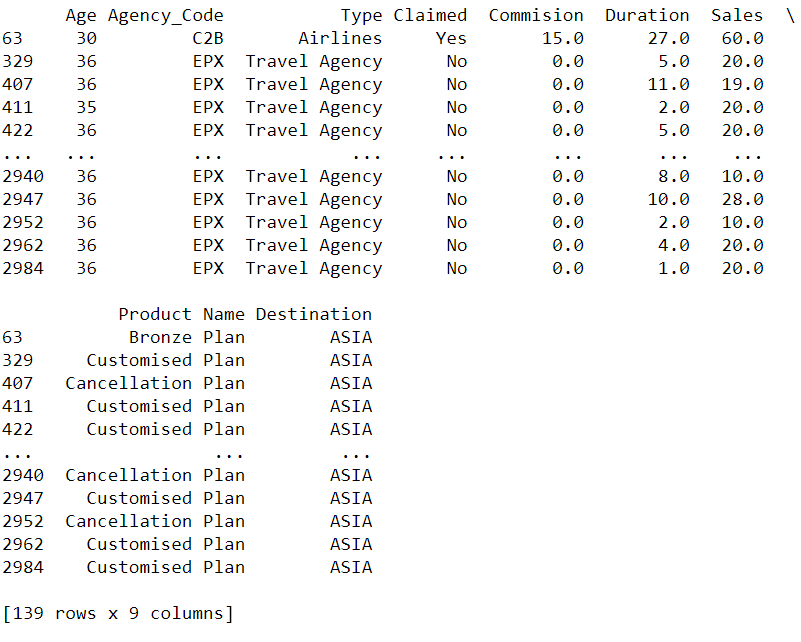


The data has 3000 rows and 10 columns.

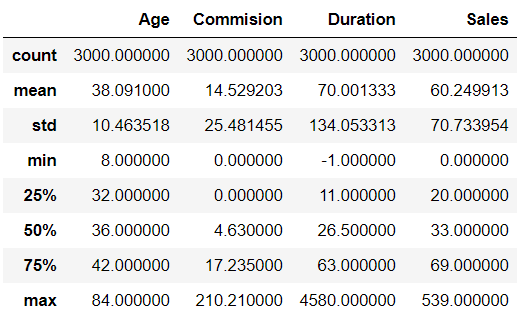
We have a combination of numeric data in some columns and object data others.



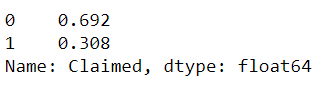
* By seeing the value counts of the variables we could observe that there was an observation as ‘-1’ in the duration column. This is obviously an incorrect entry in the data and needs to be corrected.
* There are 139 rows in the data which claim to be duplicates, but we will ignore the same An entry of contract ID number or the customer name would help us to determine if these are actual duplicate records or not. But for now, we will choose to ignore the same as it is quite possible that the same insurance company, has sold the same type of insurance to the same type customer. Below are the entries in our data that claim to be duplicates:



* There are no null entries in the data.
* We will choose the One Hot Encoding method to transform the categorical variables to numeric. The values in the categorical variables do not take any precedence over one another.
* We will choose to drop the Channel variable from the data as the Channel variable has only 46 entries belonging to Offline mode whereas the remaining have an online mode.This contributes to 1.5 % of the data and cannot have any notable impact on the target variable. Also, going forward as the digitisation is taking over manual entry, the offline mode of distribution channel for tour insurance agencies would be eliminated. Hence we will choose to drop them.
* Data distribution in the numerical columns: We can say that the data is highly right skewed. The minimum value in the duration column is shown as -1 which will be treated as mentioned before.

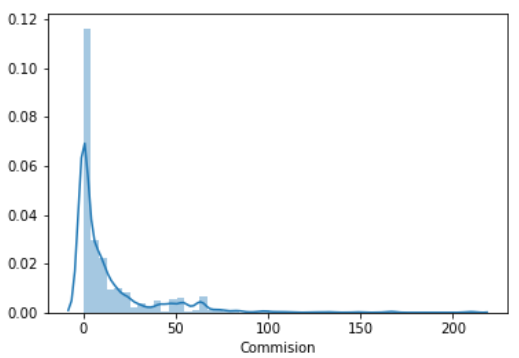


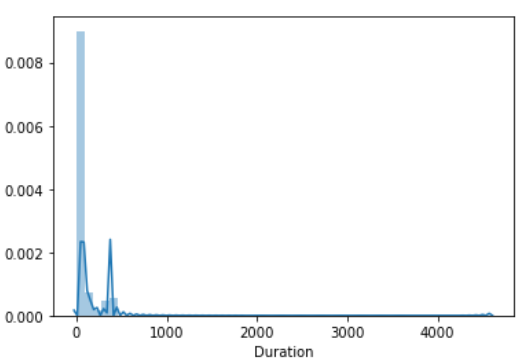
* We then look at the dirtribution of data in the target variable: Claimed

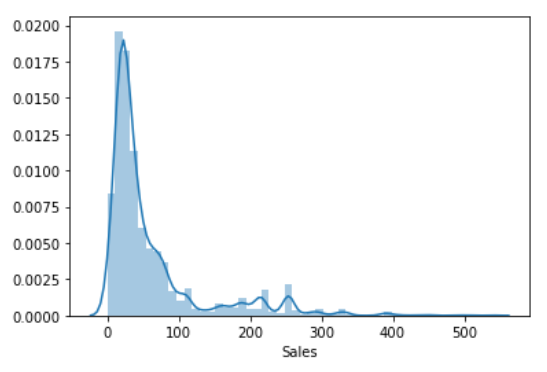


It is seen that the target variable is unequally distributed. With close to 70 percent in the 0 claim status. This means that the prediction metrics on the 0 will be more than on the 1.

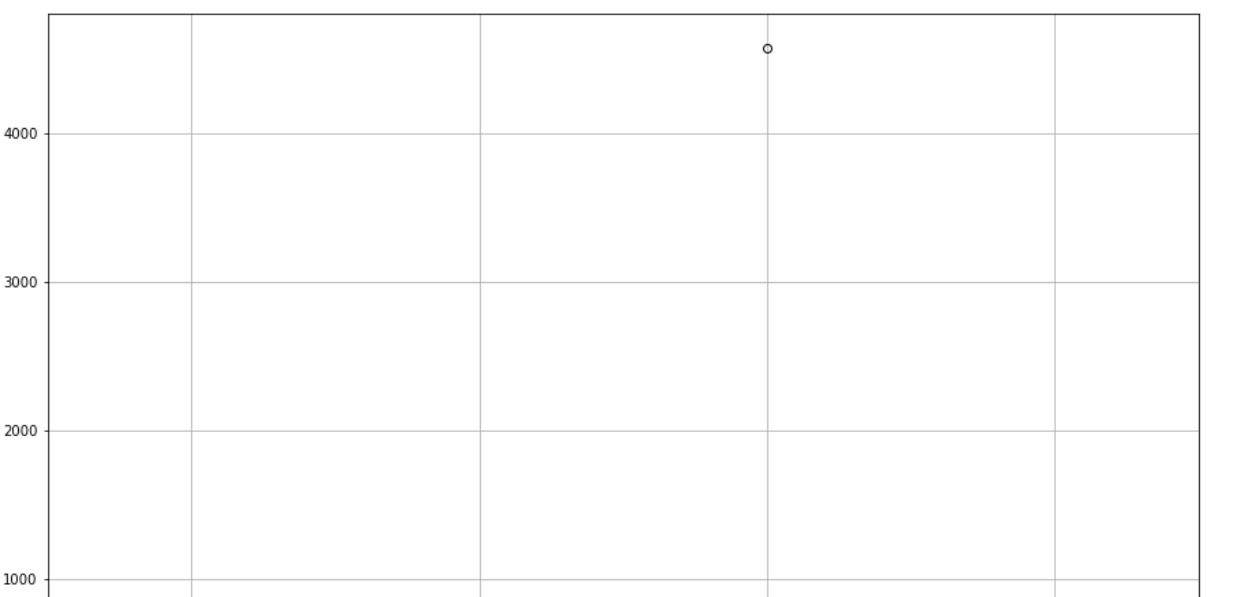
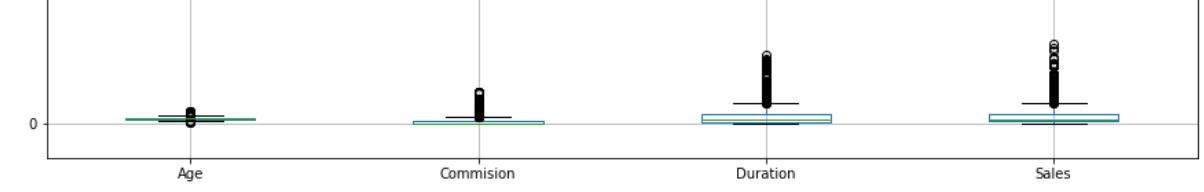
* Distribution plot on the numerical variable shows us that the data is highly right skewed.



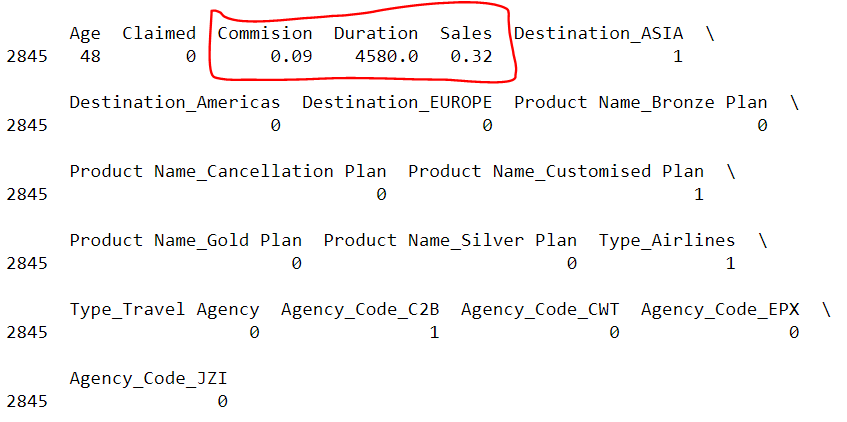




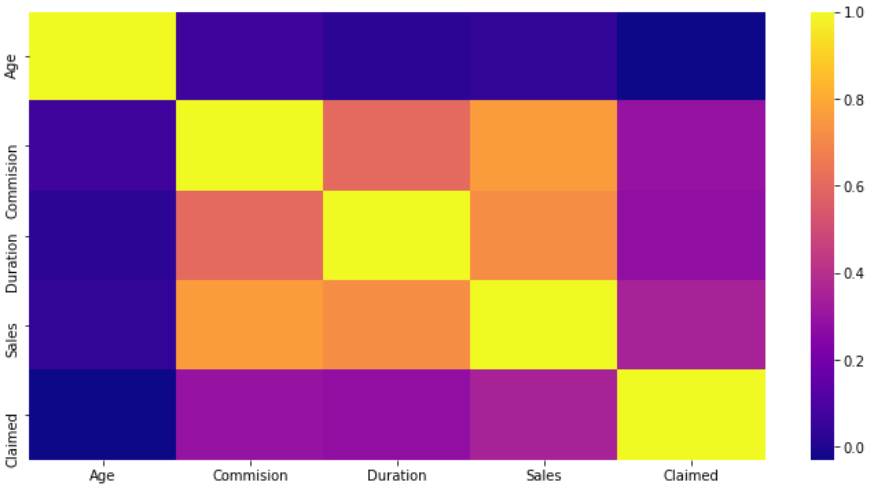
* Boxplot of the same shows us the following:

* It can be observed that there is an extremely large outlier in the duration column. Upon further investigation, it was found that it was an incorrect entry as the commission and sales figures were not in line with the duration of travel.

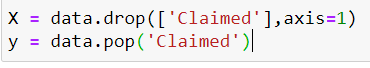


* A correlation heatmap was plotted only for the numerical variables. It can be observed that the Age variable shows very little or no relation to other numerical variables. Commision shows a high positive correlation to sales (0.8 – 1). Duration of travel also shows a significant correlation to commission and sales (0.5 – 0.8).

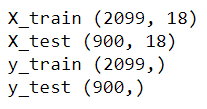


* Questions 2 to 4 will be handled in each of our trials for building the perfect model.
* **Trial 1:** Channel column has been excluded, the object variables: Destination, Product Name, Type, Agency\_Code will be treated by One Hot Encoding method, Train set size = 70% and Test set size = 30%.

Data Split and model Building:



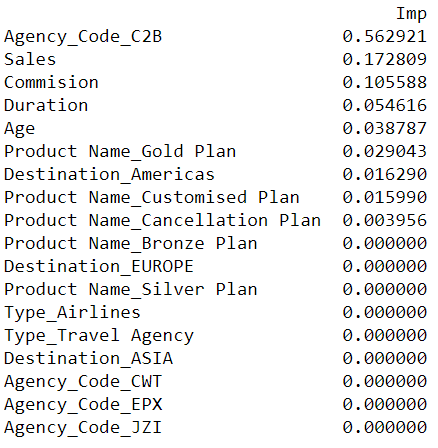




* Decision Tree Classification model was built using the Gini criterion and the Grid search cross validation method. Decision tree was also built. It was fit into the train and test data.

**Model Performance Summary:**

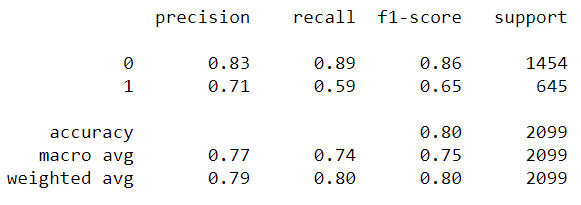
Variable Importance:



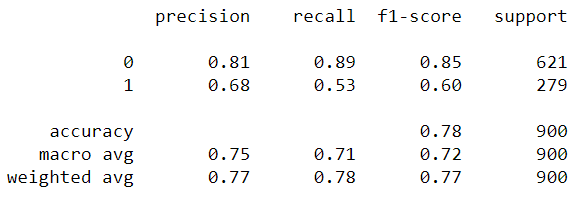
* Agency\_Code\_C2B, Sales and Commision have a high importance in predicting the outcome of the target variable: Claimed.

Classification Report:

Train Data:

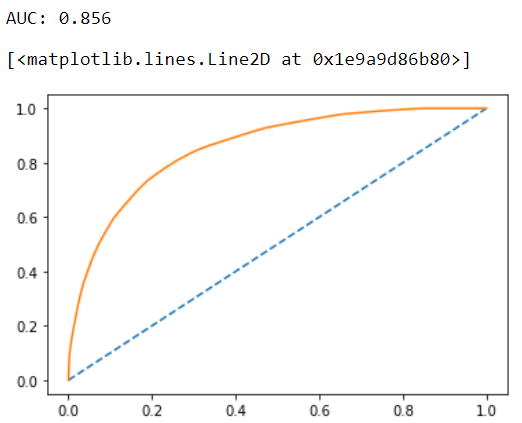


Test Data:

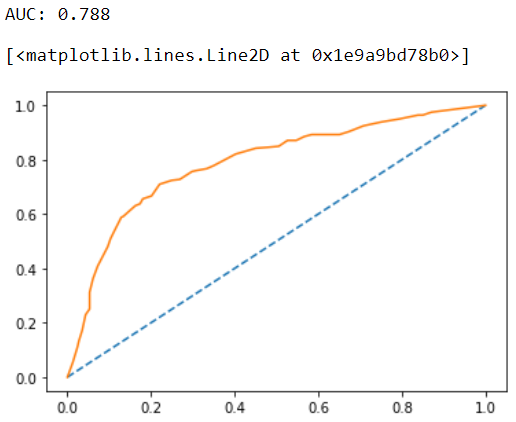


AUC & ROC:

Train Data:

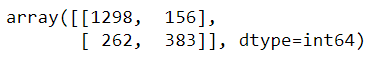


Test Data:

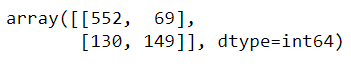


Confusion Matrix:

Train Data:



Test Data:



CART Conclusion:

|  |  |  |
| --- | --- | --- |
|  | Train Data: | Test Data: |
| AUC | 86 | 78 |
| Accuracy | 80 | 78 |
| Precision | 71 | 68 |
| F1 Score | 65 | 60 |
| Recall | 59 | 53 |
| Specificity | 89 | 89 |

The accuracy for training and test results is high. Not much difference between training and testing accuracy. Precision is low for both train ad test with only 71 and 68 respectively.

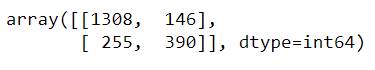
Specificity is good for both training and testing.

* Random Forest Classifier model was built using the grid search cross validation and the best grid was extracted and fit on both the train and test data.

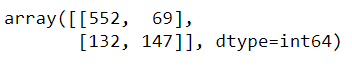
**Model Performance Evaluation**

**Confusion Matrix:**

Train data:

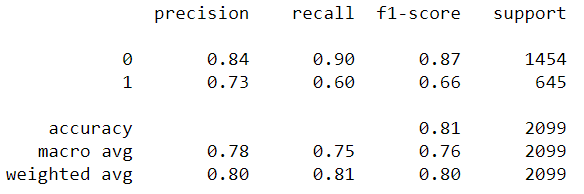


Test data:

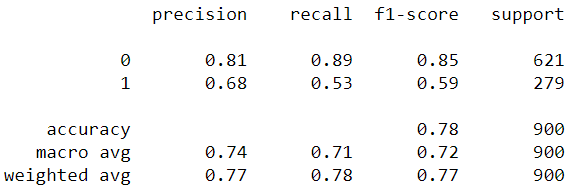


**Classification Report:**

Train data:



Test data:

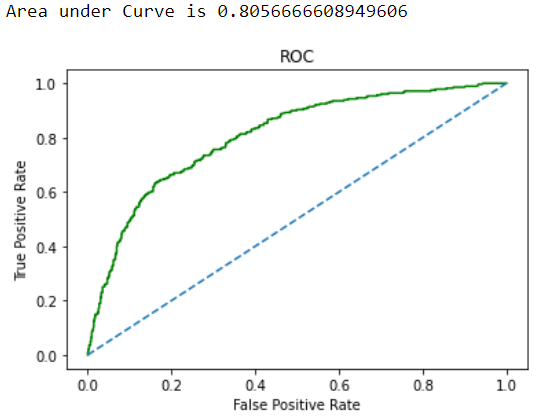


**AUC & ROC:**

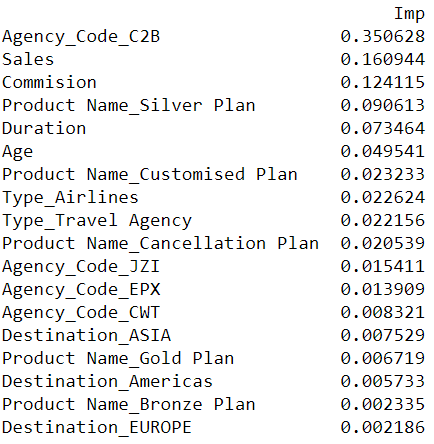
Train Data:



Test Data:



**Variable Importance:**



**Random Forest Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | **Train Data** | **Test Data** |
| AUC | 86 | 81 |
| Accuracy | 81 | 78 |
| Precision | 73 | 68 |
| F1 Score | 66 | 59 |
| Recall | 60 | 53 |
| Specificity | 90 | 89 |

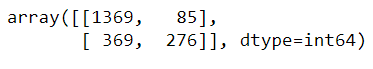
**Artificial Neural Network:**

* The Train and the Test data was scaled using the standard scaler which computes the z scores of the data.
* The MLP classifier was used along with the Grid search cross validation and the best grid was applied on the train and test data.

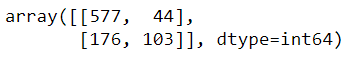
**Model Performance Evaluation:**

**Confusion Matrix:**

Train data:

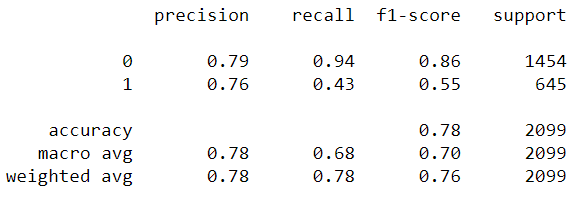


Test Data:



**Classification Report:**

Train Data:

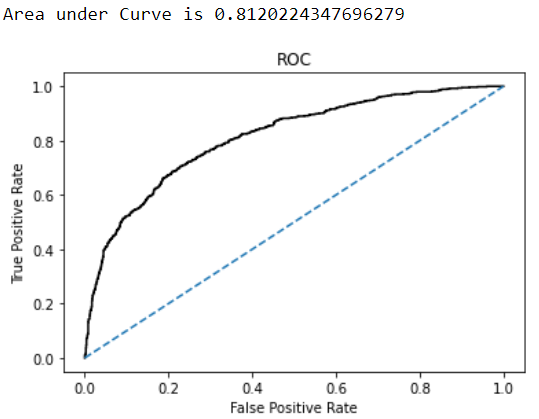


Test Data:

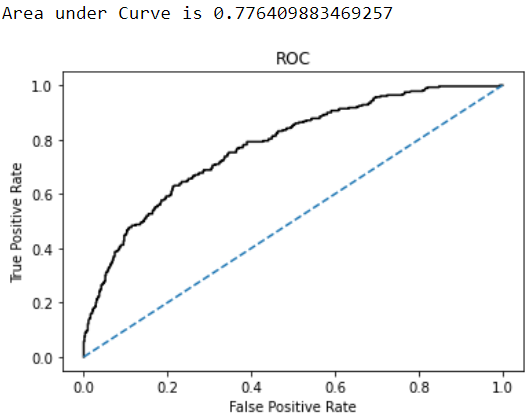


**ROC & AUC:**

Train Data:



Test data:



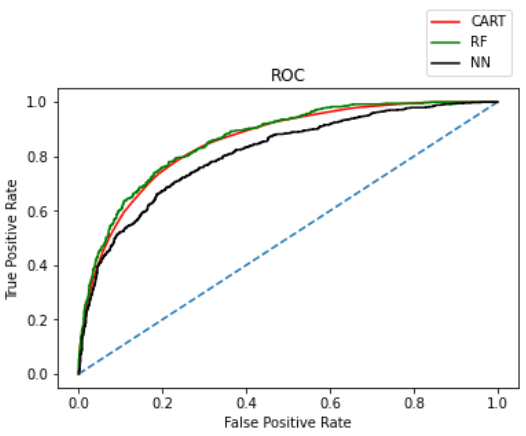
**Artificial Neural Network Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 81 | 78 |
| Accuracy | 78 | 76 |
| Precision | 76 | 70 |
| F1-Score | 55 | 48 |
| Recall | 43 | 37 |
| Specificity | 94 | 93 |

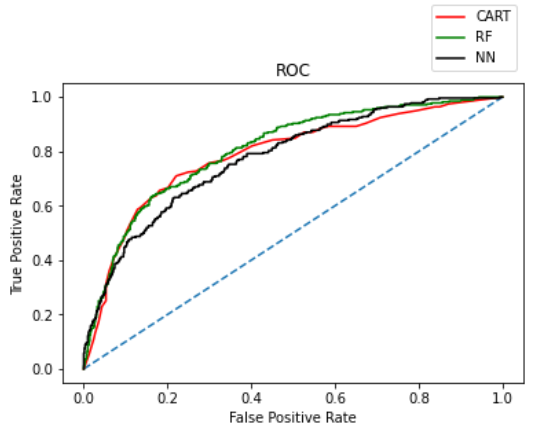
**Comparison of Performance metrics from the 3 models:**



**ROC Curve for the 3 models on the training data:**



**ROC Curve on the 3 models using the Testing Data:**



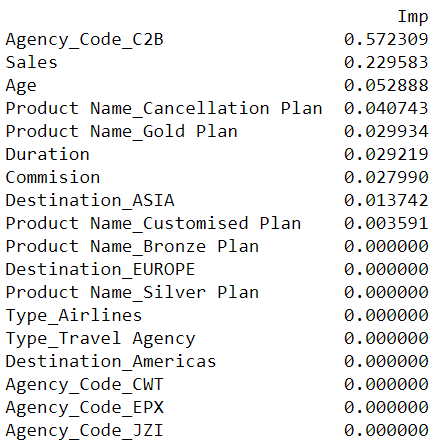
**Trial 2:**

The Train and Test sizes are changed to 80% and 20% respectively.

**Decision tree classifier:**

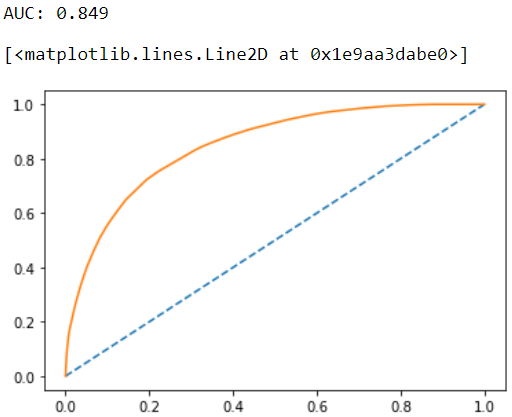
**Model Evaluation:**

**Feature Importance:**

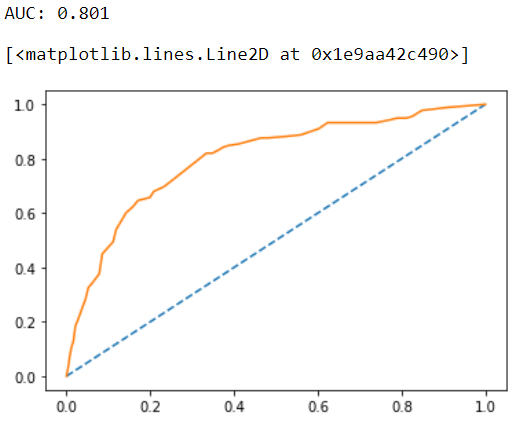


**AUC and ROC:**

Train Data:



Test Data:

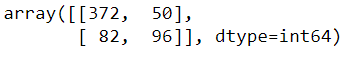


**Confusion Matrix:**

Train Data:

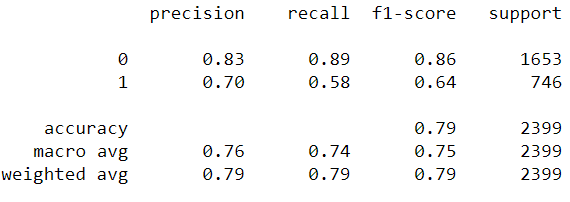


Test Data:

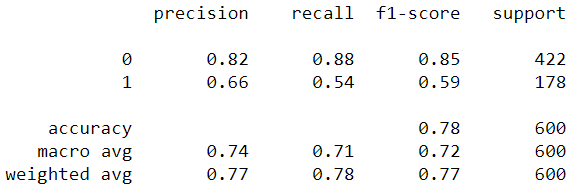


**Classification Report:**

Train Data:



Test Data:



**CART Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 85 | 80 |
| Accuracy | 79 | 78 |
| Precision | 70 | 66 |
| F1-Score | 64 | 59 |
| Recall | 58 | 54 |
| Specificity | 89 | 88 |

The training and test results for all the parameters have reduced as compared to the previous instance where we had 70% training and 30% testing.

**Random Forest Classifier:**

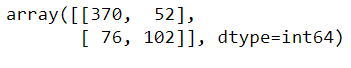
**Model performance evaluation:**

**Confusion Matrix:**

Train data:

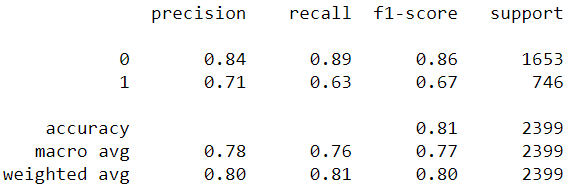


Test Data:

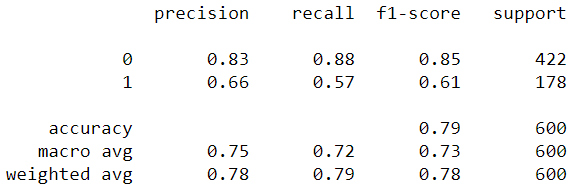


**Classification report:**

Train Data:

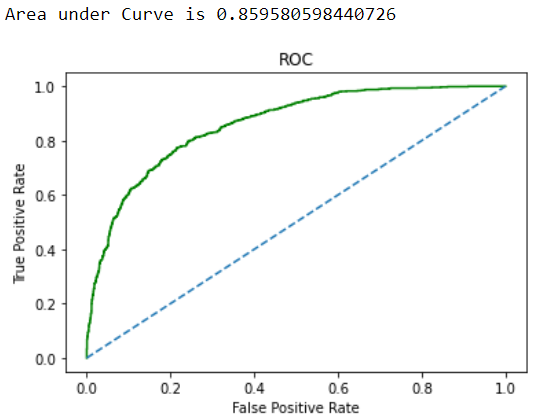


Test Data:

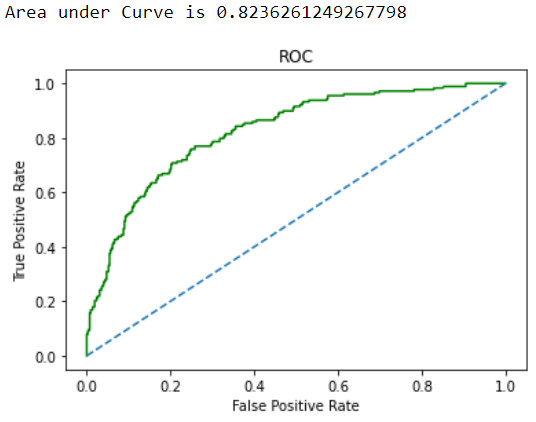


**ROC & AUC:**

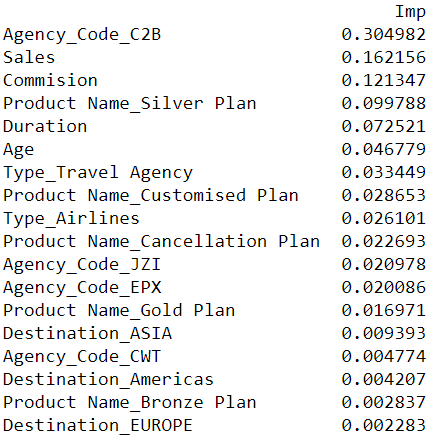
Train data:



Test data:



**Variable Importance:**



**Model performance summary:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 86 | 82 |
| Accuracy | 81 | 79 |
| Precision | 71 | 66 |
| F1 Score | 67 | 61 |
| Recall | 63 | 57 |
| Specificity | 89 | 88 |

**Artificial Neural Network:**

**Model Performance Evaluation:**

**Confusion Matrix:**

Train data:

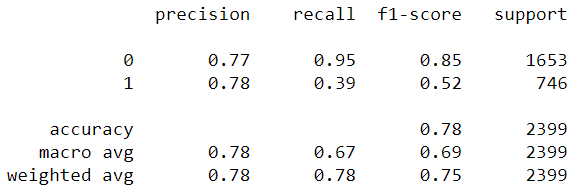


Test Data:

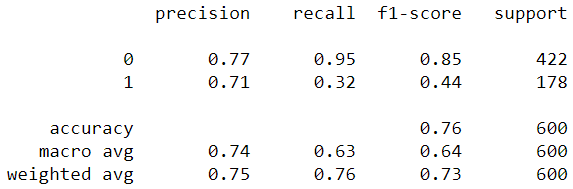


**Classification Report:**

Train Data:

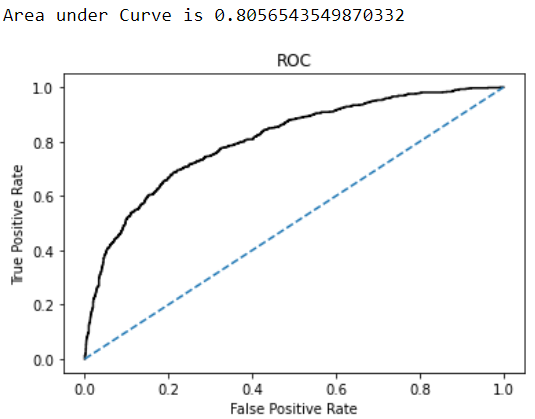


Test data:

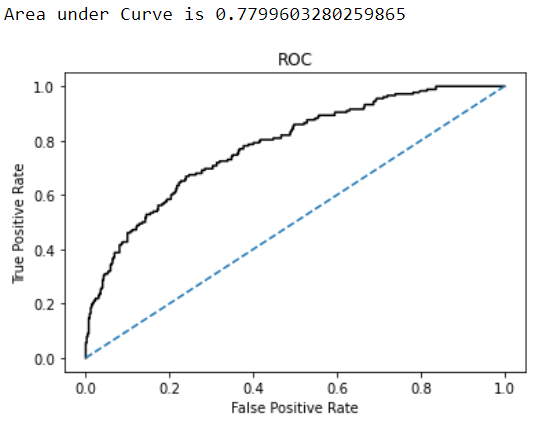


**ROC & AUC:**

Train data:



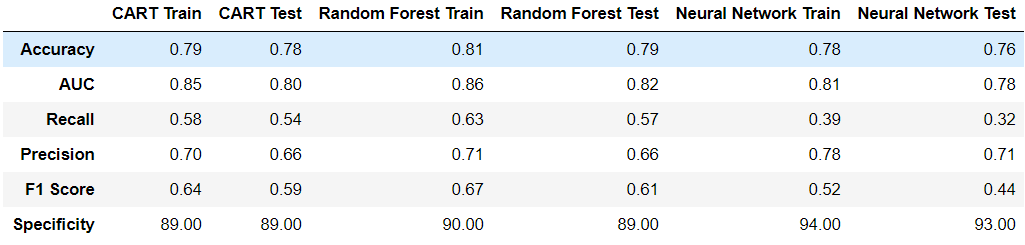
Test data:



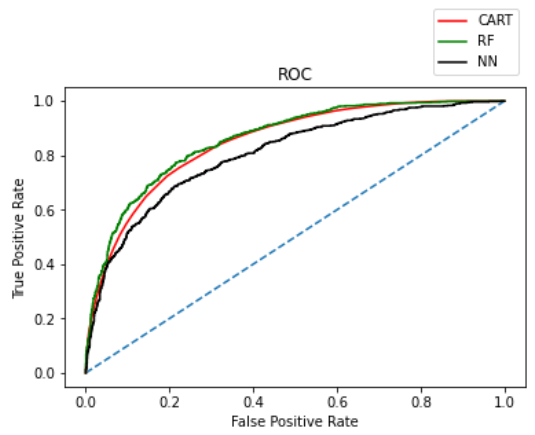
**Artificial Neural Network Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 81 | 78 |
| Accuracy | 78 | 76 |
| Precision | 78 | 71 |
| F1 Score | 52 | 44 |
| Recall | 39 | 32 |
| Specificity | 95 | 95 |

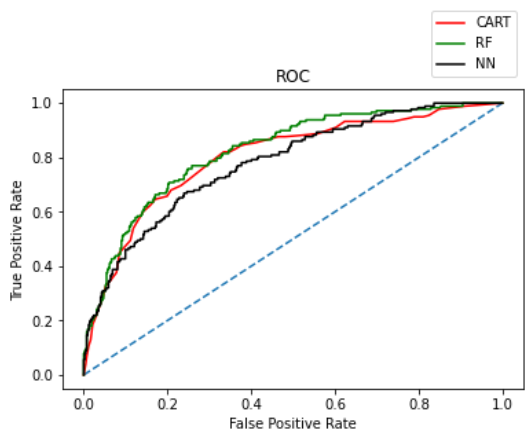
**Comparison of Performance metrics from the 3 models:**



**ROC Curve for the 3 models on the training data:**



**ROC Curve on the 3 models using the Testing Data:**

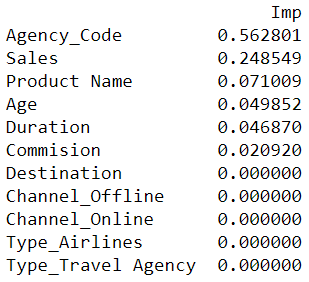


**Trial 3:** In this we will include all the variables and use label encoding method for certain categorical variables and one hot encoding for other categorical variables. The incorrect value in the duration column has also been replaced with the median. Train and Test sizes are 70% and 30% respectively.

**Decision tree classifier:**

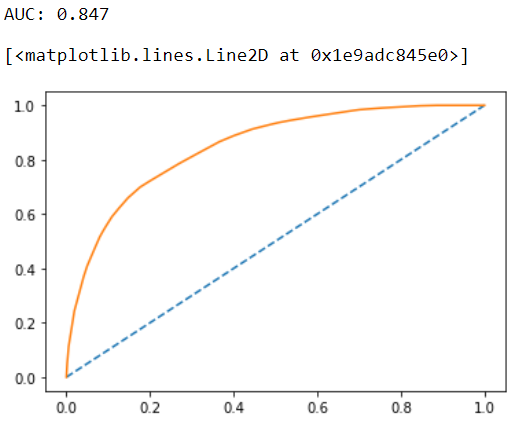
**Model Evaluation:**

**Feature Importance:**

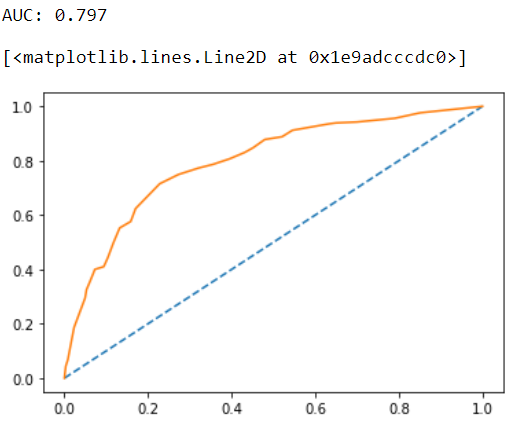


**AUC & ROC for train and test data:**

Train data:



Test data:

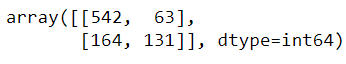


**Confusion Matrix:**

Train data:

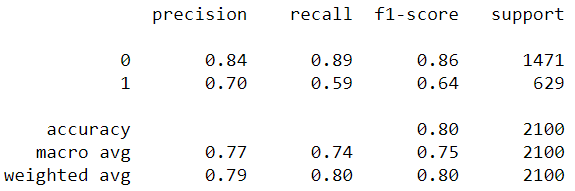


Test data:

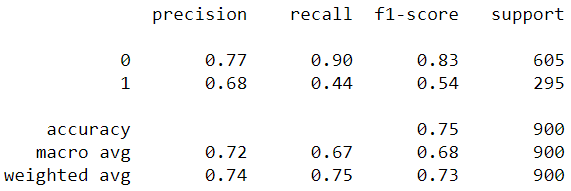


**Classification Report:**

Train data:



Test data:



**Accuracy:**

Train data: 80.09%

Test data: 75%

**CART Conclusion:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 85 | 80 |
| Accuracy | 80 | 75 |
| Precision | 70 | 68 |
| F1-Score | 64 | 54 |
| Recall | 59 | 44 |
| Specificity | 89 | 90 |

The Precision, F1 score and Recall values on the test data have improved as compared to our previous CART trial.

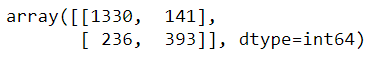
However, the accuracy and specificity have actually gone down on this trial.

**Random Forest Classifier:**

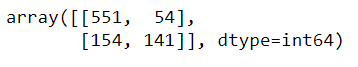
**Model Evaluation:**

**Confusion Matrix:**

Train data:



Test data:



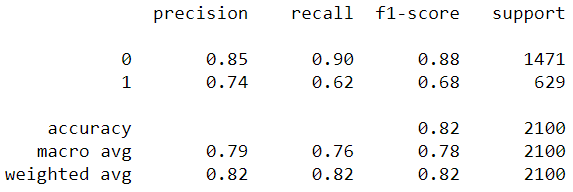
**Accuracy:**

Accuracy for the train data on RF model is 82.04%

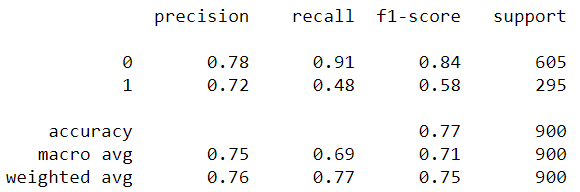
Accuracy for the test data on RF model is 76.8%

**Classification Report:**

Train data:

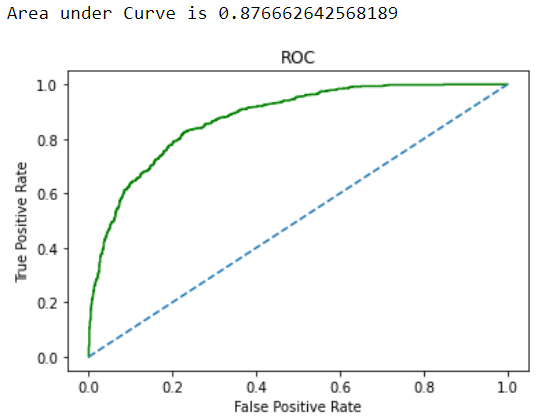


Test data:

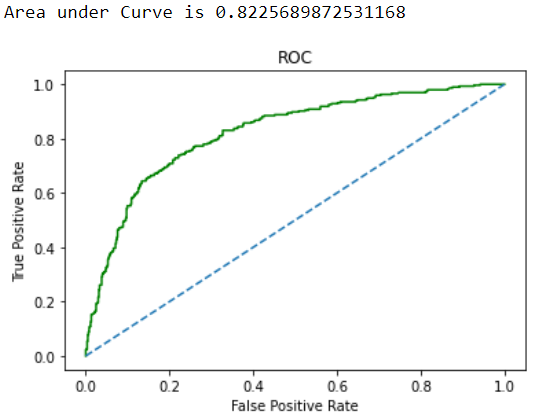


**ROC AUC Curve:**

Train data:



Test data:



**Variable Importance:**



**Model performance summary:**

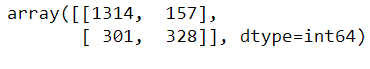
|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 88 | 82 |
| Accuracy | 82 | 77 |
| Precision | 74 | 72 |
| F1-Score | 68 | 58 |
| Recall | 62 | 48 |
| Specificity | 90 | 91 |

**Artificial Neural Network:**

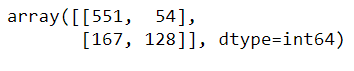
**Model Evaluation:**

**Confusion Matrix:**

Train:



Test:



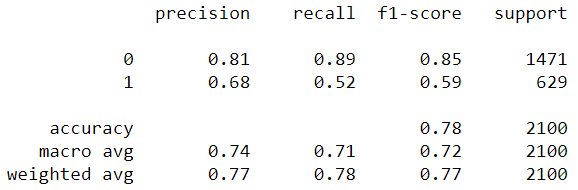
**Accuracy:**

Accuracy on the train data is 78.19%

Accuracy on the test data is 75.44%

**Classification Report:**

Train data:

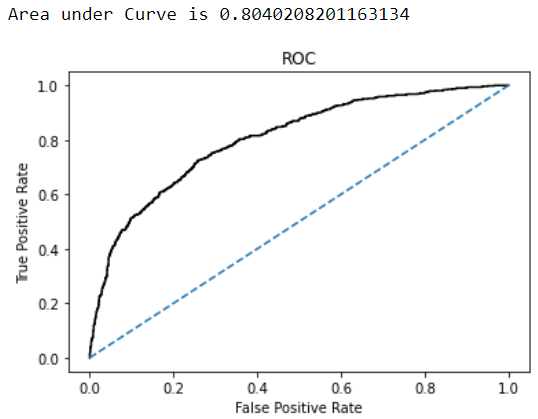


Test data:

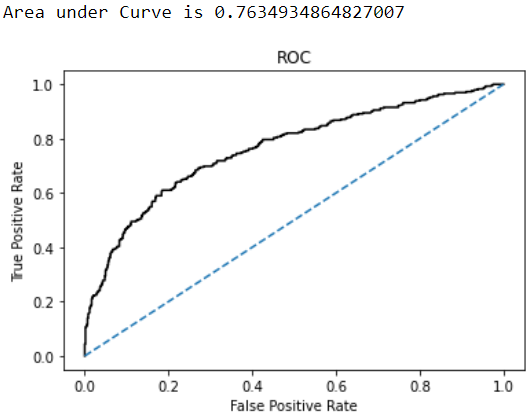


**AUC and ROC:**

Train data:



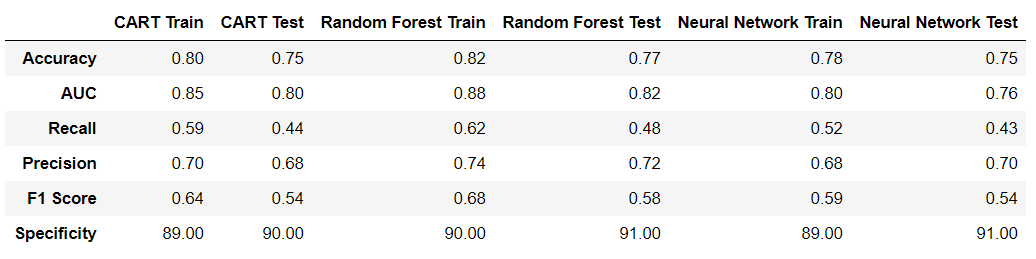
Test data:



**Model Performance Summary:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 80 | 76 |
| Accuracy | 78 | 75 |
| Precision | 68 | 70 |
| F1-Score | 59 | 54 |
| Recall | 52 | 43 |
| Specificity | 89 | 91 |

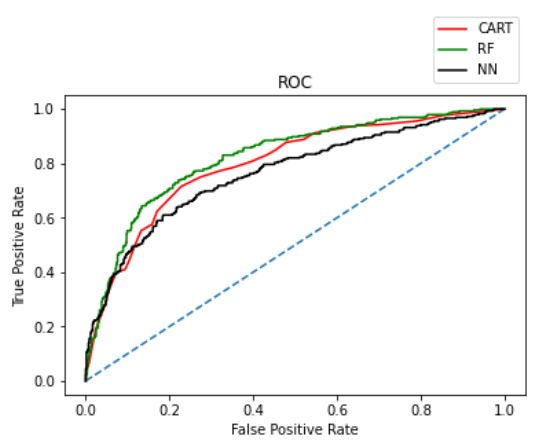
**Comparison of the performance metrics from the 3 models:**



**ROC Curve for the 3 models on the training data:**



**ROC Curve on the 3 models using the Testing Data:**



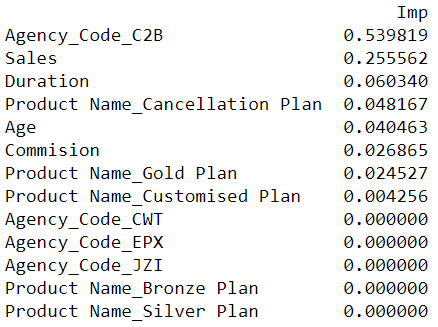
**Trial 4:**

In this trial, we will remove the variables: Destination, Channel, Type from our analysis. This is because these variables have consistently proven to be least important among all the various trials we have performed so far with the model building of the data. The train and test sizes have been split into 70% and 30% respectively.

**Decision Tree Classifier:**

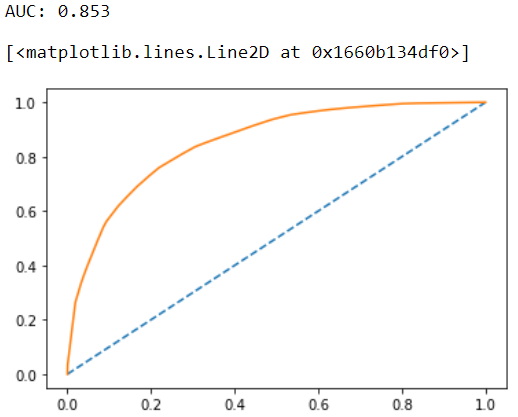
**Model Evaluation:**

**Variable Importance:**

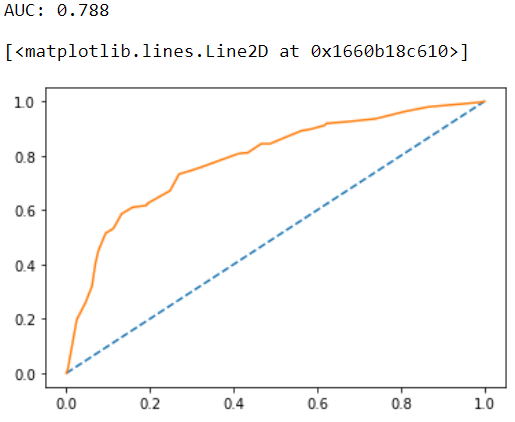


**AUC & ROC:**

Train data:



Test data:

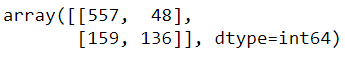


**Confusion Matrix:**

Train data:

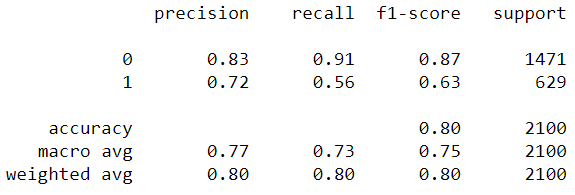


Test data:

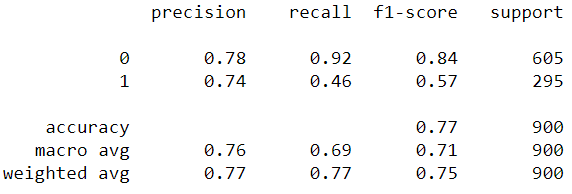


**Classification Report:**

Train data:



Test data:



**Accuracy:**

Accuracy of the train data is 80.33%

Accuracy of the test data is 77%

**CART Model Performance Summary:**

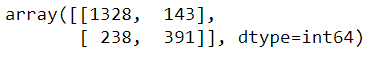
|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 85 | 79 |
| Accuracy | 80 | 77 |
| Precision | 72 | 74 |
| F1-Score | 63 | 57 |
| Recall | 56 | 46 |
| Specificity | 91 | 92 |

**Random Forest Classifier:**

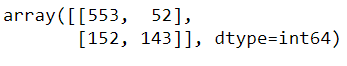
**Model Evaluation:**

**Confusion Matrix:**

Train data:



Test data:



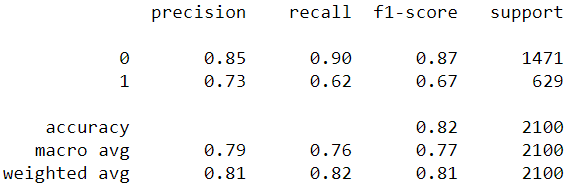
**Accuracy:**

Accuracy for the Train data on RF model is 81.85%

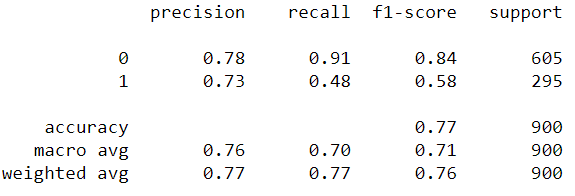
Accuracy for the Test data on the RF model is 77.33%

**Classification Report:**

Train data:

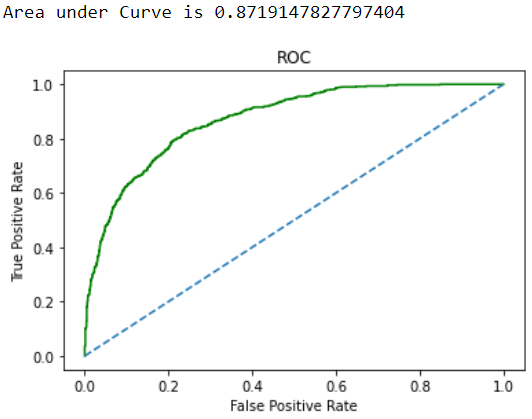


Test data:

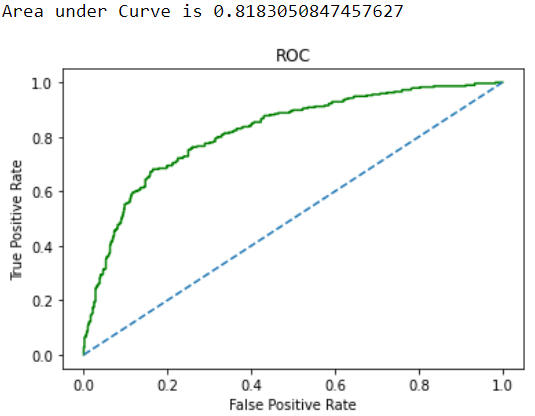


**AUC & ROC:**

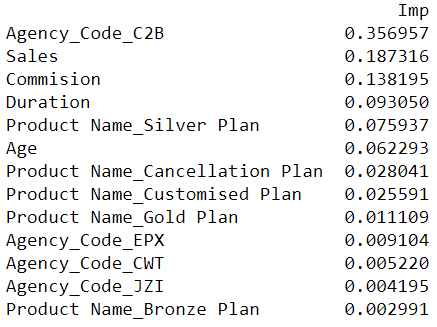
Train data:



Test data:



**Variable Importance:**



**Random Forest Model Performance Summary:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 87 | 82 |
| Accuracy | 82 | 77 |
| Precision | 73 | 73 |
| F1-Score | 67 | 58 |
| Recall | 62 | 48 |
| Specificity | 90 | 91 |

**Artificial Neural Network:**

**Model Performance Evaluation:**

**Confusion Matrix:**

Train data:



Test data:



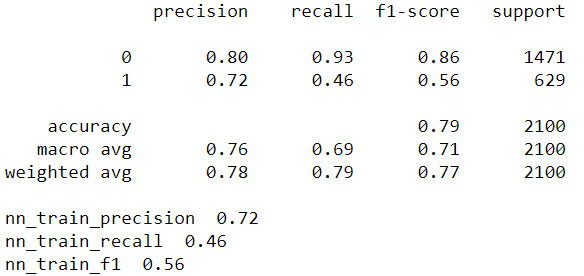
**Accuracy:**

Accuracy for the train data is 78.52%

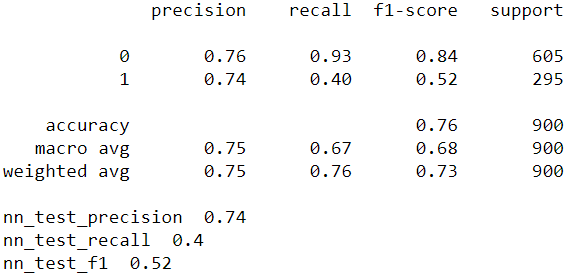
Accuracy for the test data is 75.77%

**Classification Report:**

Train data:

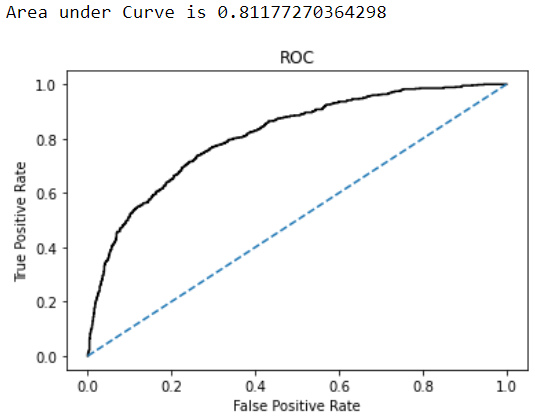


Test data:

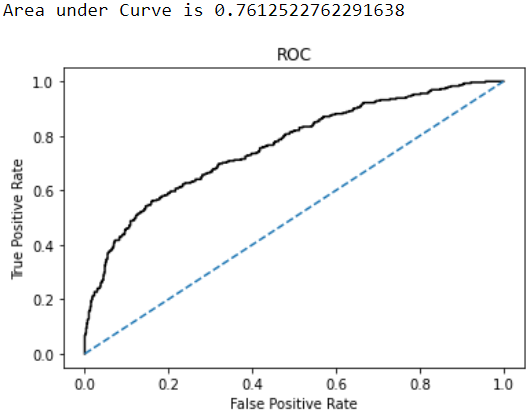


**AUC and ROC:**

Train data:



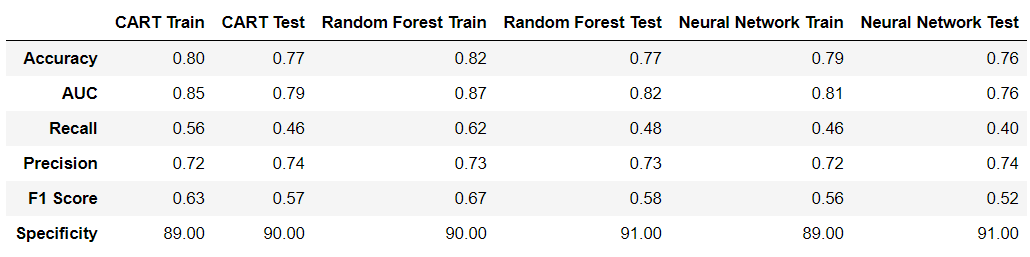
Test data:



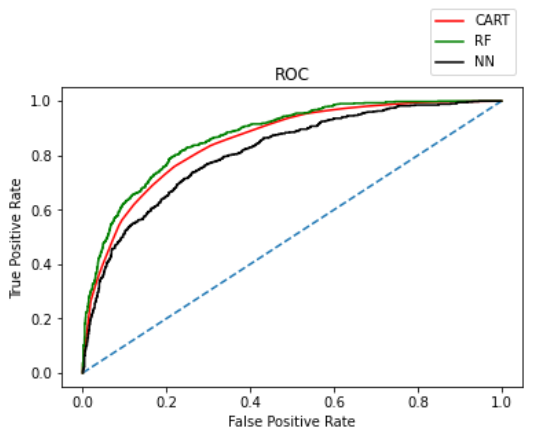
**ANN Model performance summary:**

|  |  |  |
| --- | --- | --- |
|  | Train | Test |
| AUC | 81 | 76 |
| Accuracy | 78 | 76 |
| Precision | 72 | 74 |
| F1-Score | 56 | 52 |
| Recall | 46 | 40 |
| Specificity | 93 | 93 |

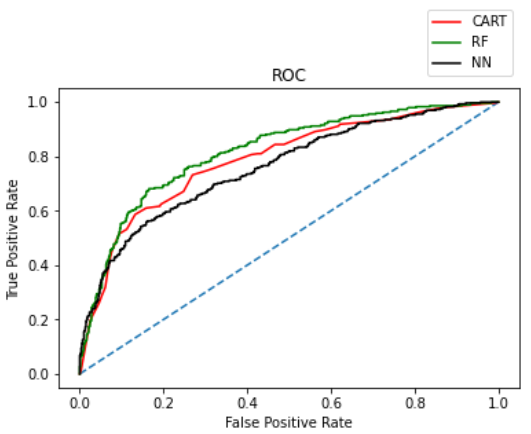
**Comparison of performance metrics from the 3 models:**



**ROC Curve for the 3 models on the training data:**



**ROC Curve on the 3 models on the Testing Data:**



**Inferences:**

* We will not look at the accuracy scores of the model in any of the trials as the values are not evenly distributed in the target variable: Claimed. We have 70% of 0 and 30% of 1. We cannot rely on this metric when the class representation is lopsided as algorithms are biased towards over represented classes.
* Precision is an important metric in our analysis and it is highest for the test data in the Neural Networks model of our Trial 4 (where we removed the poorly performing variables: Destination, Channel, Type). Precision for test data is 74% and is better than that of the train data which is 72%.
* Recall is consistently low in each of our versions of analysis and it is highest for our Random Forest Model in the trial 2 (Channel variable is dropped and train-test size is 80-20 respectively). Recall for train data is 63% and for test data is 57%.
* The Random Forest model has consistently proven to be the better model in the ROC & AUC curve in all the trials. The AUC highest for the test data is 82% and it is for the Random Forest model.
* In our business problem, we have been given to design a model to predict the claim status of tour insurance. This means that it is important that we correctly identify the 0 – Not Claimed. Specificity is a metric that is the one of the most important metrics in our analysis and it is highest for the test data in the ANN model for both trials 1 and 2. It is also to observe that the only difference between the trials 1 and 2 is the difference in the train and test size. Also, as previously mentioned, the data in the target column: Claimed is not even. This means that a model can learn better to predict the 0 than to predict the 1 as there is more data available for 0s where the model can learn. Among the trials 1 and 2, the True Negatives is the highest for the test data in the ANN model of Trial 1 – 577.
* It is extremely important for the model to have the lowest False Negatives – cases where the actual class was positive (1) but was predicted as negative (0). This is where the insurance firm can benefit the most from the model implementation as more number of claims means more money being spent out by the firm. Insurance company would want to be able to predict correctly beforehand if the customer is going to claim or not. And having a model where the customer is predicted as not going to claim and he ends up claiming insurance is going to mean that money is going out of the insurance firm.
* In our analysis, the False Negatives of the various models are listed as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | CART | RF | ANN |
| Trial 1 | 130 | 132 | 176 |
| Trial 2 | 82 | 76 | 121 |
| Trial 3 | 164 | 154 | 167 |
| Trial 4 | 159 | 152 | 176 |

Random Forest model in the trial 2 has the lowest False Negatives.

* Agency\_Code\_C2B has been most important in the affecting the claim status followed by sales. This agency has to be analysed specially and we need to find out the principal factors contributing to its very high influence on the insurance claim status.