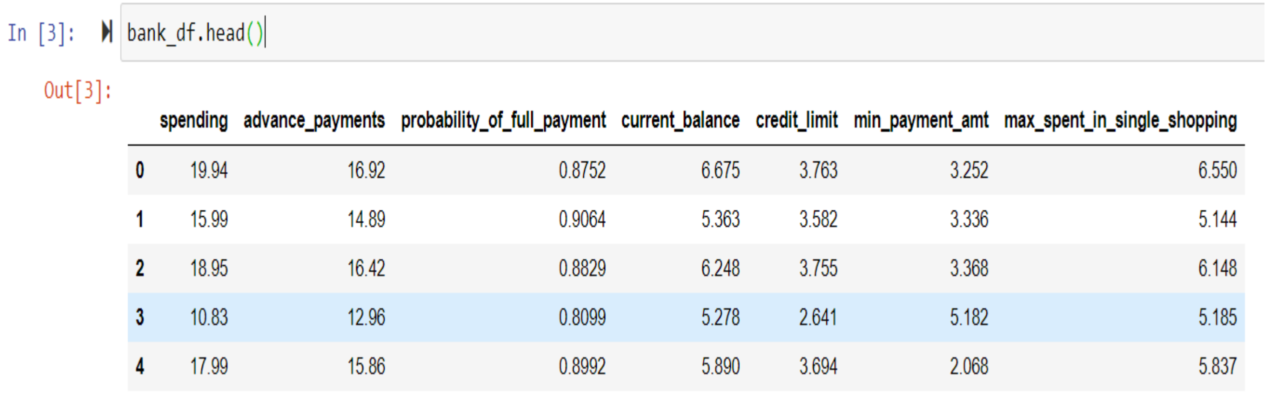
Project Topic: Data Mining

Student Name: Ajay Patwari

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 and do exploratory data analysis. Describe the data briefly.

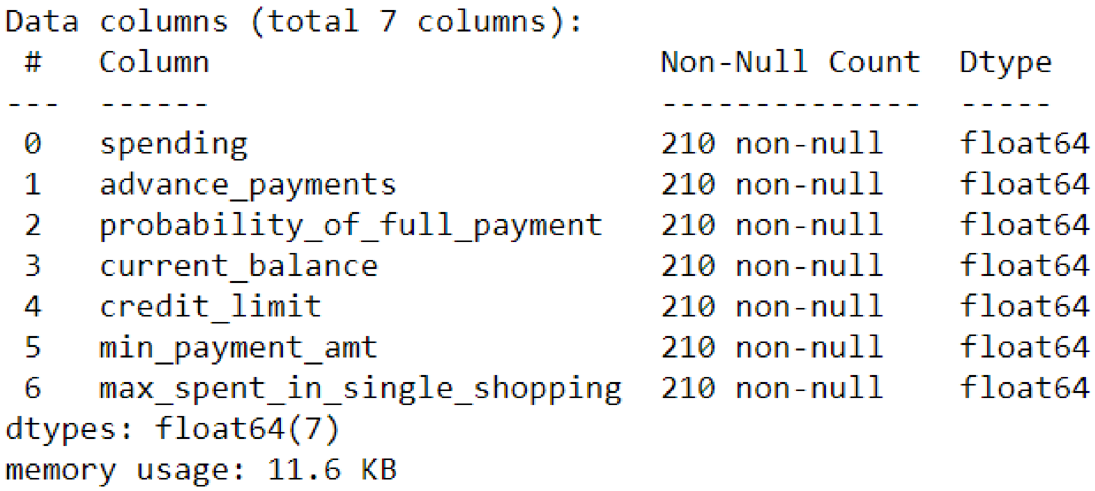


```
In [3]:  bank_df.head()
```

Out[3]:

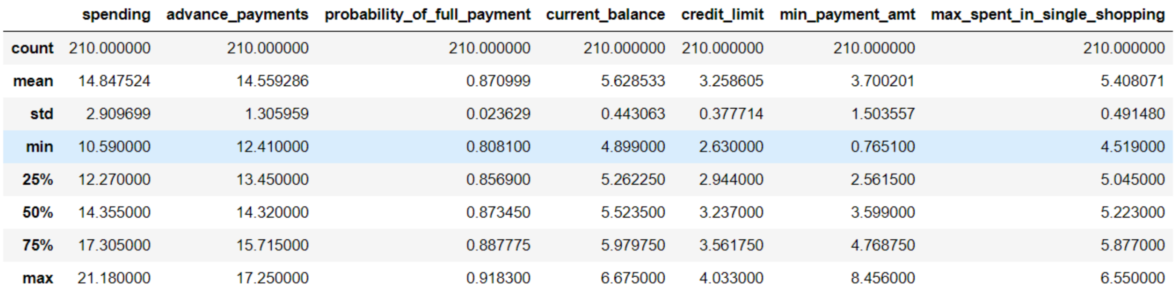
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

The above is the figure of the first 5 rows of the data set given. We can see that there are 7 columns. The total number of columns of this data set is 210.



```
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

We can see that all the columns of the given data are of Float Data Type.



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

The Below figure helps us describe the whole data and gives us the mean, standard deviation, minimum, maximum values of the data.

We have used the duplicated and isnull functions to confirm that there are no duplicated values and no missing values in the given data.

From the given data it can be inferred that the probability of full payment of the customers is on an average 87% and if you take the median also the value does not vary much it is almost similar to the mean.

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

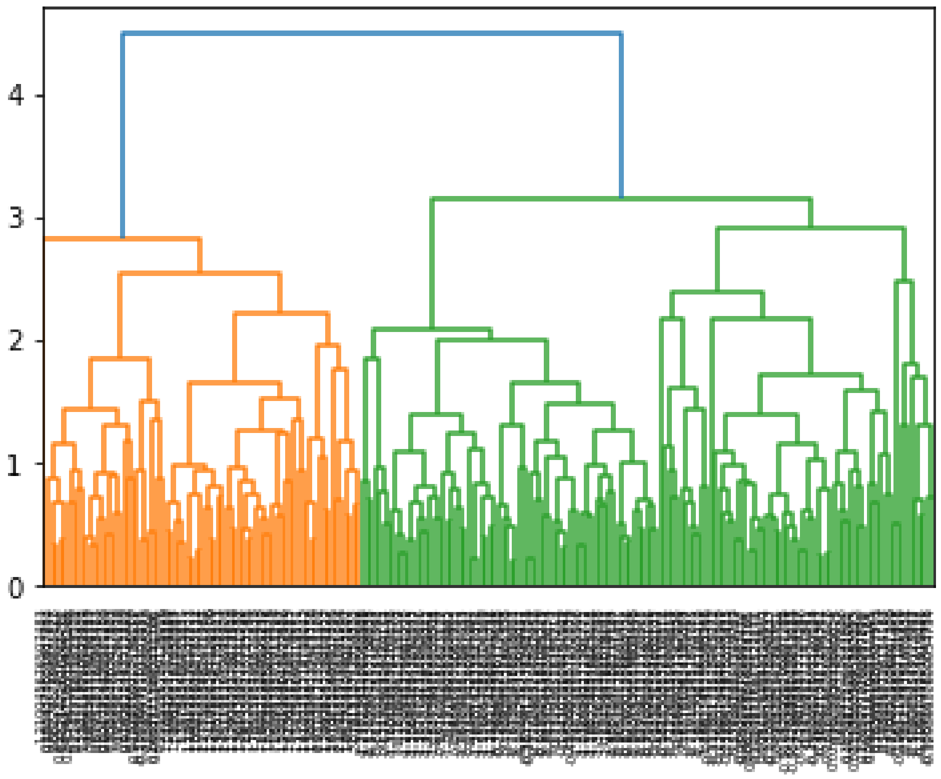
Yes, scaling for the given data is mandatory because the given data are all of different values. The spending's, the current balance, max spent in a single shopping are given in 1000s value, advance payments are given in 100s value and the credit limit is given in 10000s value.

This means that there are discrepancies with the data and we must make sure that the data is scaled and made sure that all the values are on the same scale.

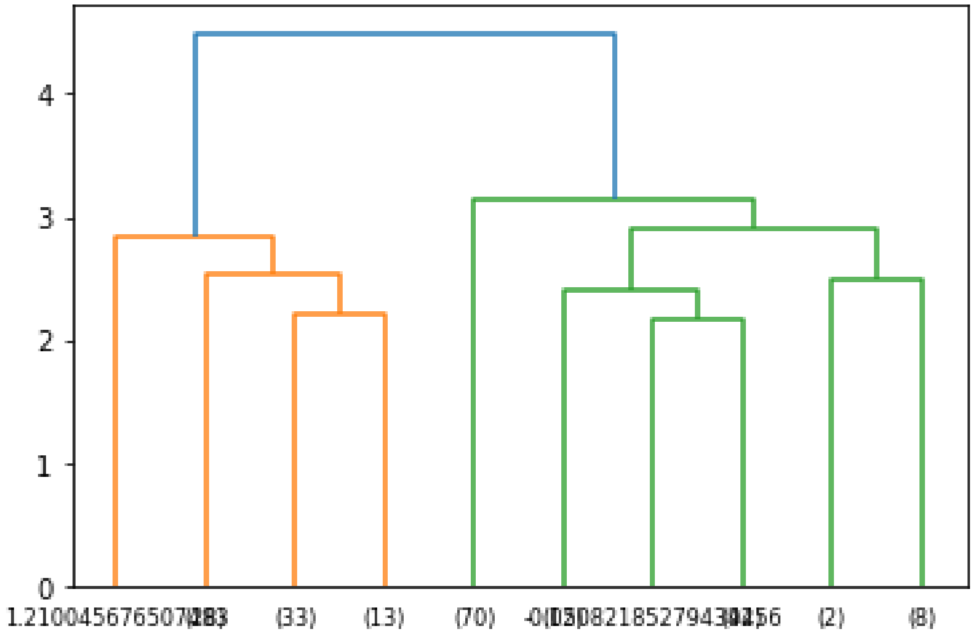
Hence, I've used the Standard Scaler method and done the scaling part for the data.

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

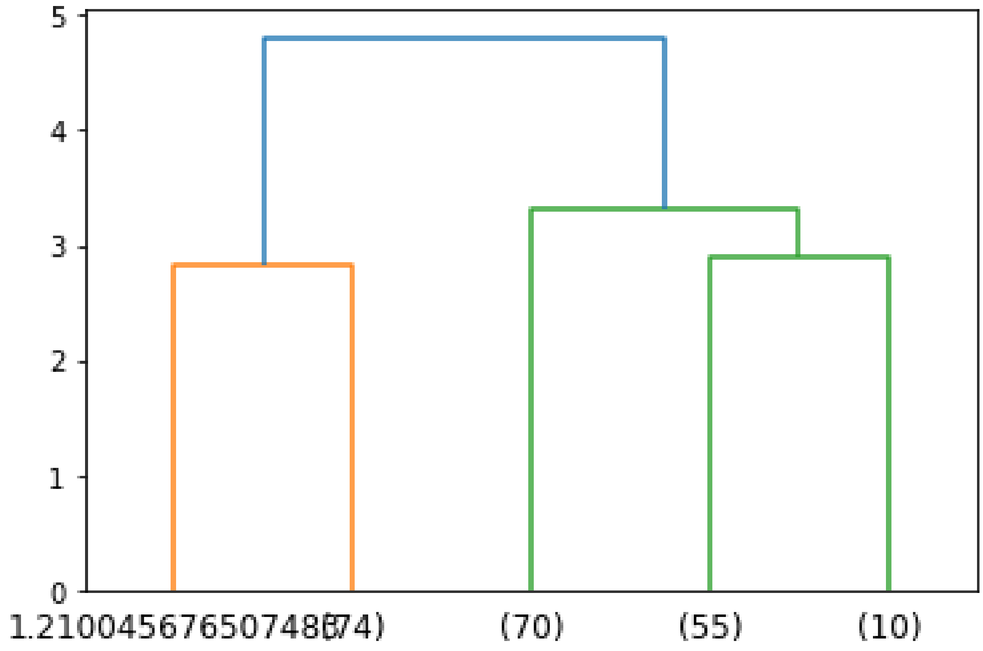
Let me first explain about the Dendrogram. A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.



This is the first dendrogram formed with the maximum number of clusters possible.



The above is the dendrogram diagram taken with the last 10 clusters. We can clearly see that we do not see much differences after the 3rd clusters.

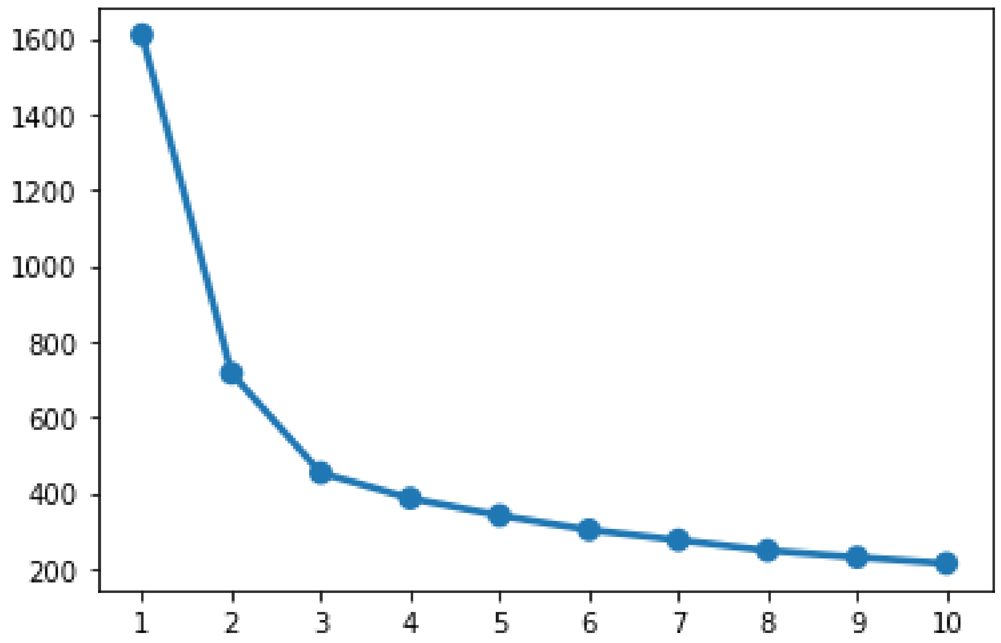


To make it more clear, I've only taken the last 5 clusters formed that the data and we can clearly come to a conclusion that after 3 clusters there is not much differences and we can identify that 3 clusters would be good and the best option for the given data.

Hence, I would like to conclude that 3 clusters would be the optimum clusters using Dendrogram.

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

K means Clustering has been done on the scaled data and all the coding part has been attached in the Jupyter Notebook. Below are the elbow curve graph and silhouette score.



From the above elbow graph we can clearly see that when the cluster are only 1 the within Cluster Sums of Squares (WSS) is as high as 1600, however when two clusters are formed the WSS decreases up to 716 and when three clusters are formed the WSS is 452. However, if you look at the graph we can clearly see that after 3 clusters there isn't much difference in the WSS when compared to the results we have seen earlier. Hence is can also be inferred that 3 clusters would be a good number for creating clusters for the given data.

Now I've taken silhouette score for consideration and below are the results obtained when K values has been taken as 3,4 and 5.

When the K value is equal to 3 the silhouette score is 0.44

When the K value is equal to 4 the silhouette score is 0.38

When the K value is equal to 5 the silhouette score is 0.32

Hence, as the silhouette score is better for 3 clusters than for 4 and 5 clusters. So, final clusters will be 3.

The procedure used for deriving these results are clearly shown in the codebook.

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

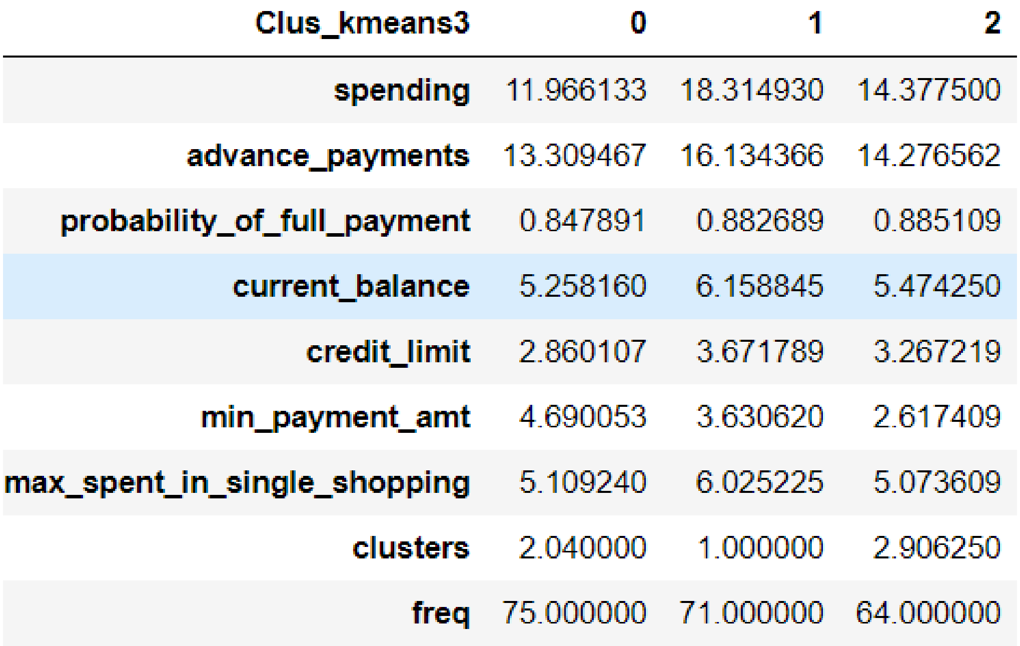
After the process of K-means clustering is done, we have got the below output when 3 clusters are made which is the best number of clusters for the given data.

The three clusters divide data and form clusters where:

Cluster 0 has 75 rows

Cluster 1 has 71 rows

Cluster 2 has 64 rows



Clus_kmeans3	0	1	2
spending	11.966133	18.314930	14.377500
advance_payments	13.309467	16.134366	14.276562
probability_of_full_payment	0.847891	0.882689	0.885109
current_balance	5.258160	6.158845	5.474250
credit_limit	2.860107	3.671789	3.267219
min_payment_amt	4.690053	3.630620	2.617409
max_spent_in_single_shopping	5.109240	6.025225	5.073609
clusters	2.040000	1.000000	2.906250
freq	75.000000	71.000000	64.000000

The below promotional strategies can be used for different clusters:

1) If we look at the cluster 0, we can see that even though the credit limit and current balance is lesser than the customers of cluster 2, the maximum spent on the single shopping is higher for cluster 0 when compared to cluster 2 also we can see that the balance of the account is less for cluster 0 customers. Hence, we can promote the customer with increasing their credit limit which will might make them do more purchases. Also, the cluster 0 has the highest value for minimum payment amount which is very good for the customer because he will be paying minimum amount and the bank can gain interest amount from it.

2) From the cluster 1, we can easily identify that the spending's are highest for the customers in this cluster. Also, we can see that the advance payments paid by the customers in the form of cash is highest. Hence the best strategy for this cluster of customers is to give more promo offers for advance payments in cash. Let me explain it to you why by simply explaining the process of how this cash advance works and how it is beneficial to the banks.

How a cash advance works?

If you carry only credit cards for day-to-day spending, you could find yourself in a pinch when confronted with a cash-only situation, such as buying lunch from a street vendor, veggies at a farmer’s market or a sandwich at a mom-and-pop deli. In that case, a cash advance might be tempting. Some people also turn to credit card cash advances when they need paper money but don't have enough in their bank account.

Getting a cash advance is easy, but it's one the costliest ways to get your hands on some cash. This is because cash advances can come with a variety of expenses:

a) Cash advance fees. These are imposed by your card issuer. Some cards charge a flat fee per cash advance, say $5 or $10. Others charge a percentage of the amount advanced — often as much as 5%. Sometimes it's a percentage with a minimum dollar amount — such as 3% or $10, whichever is greater.

b) ATM or bank fees. These are imposed by the financial institution that handles the transaction — the owner of the ATM or the bank where you get your advance.

c) Interest. This can be costly in two ways. First, the interest rate that a credit card charges on cash advances is often much higher than the rate charged on purchases. Second, interest on cash advances usually starts accruing immediately. There's no grace period like you can get with purchases.

Hence this way it is very much beneficial for the bank for the customers to make cash advance payments which these cluster of customers are already used to. Hence promotional offers on cash advance would be the best bet.

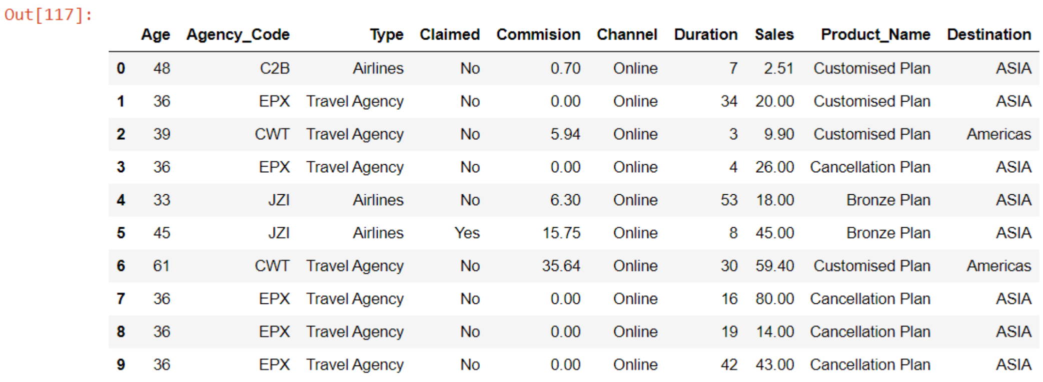
3) The cluster 2 customers are having a high probability that the payments will be done in full. These cluster of customer has the lowest number when you look at the minimum amount payment, this means the customer more often than not clears the complete amount of his credit card which makes him a genuine customer. We should offer increase in his credit limit and provide promotions when high amount is spent on a single shopping. This way it would be beneficial for the banks.

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

Below is the tablet when the head function is given for top 10 rows.



Out[117]:

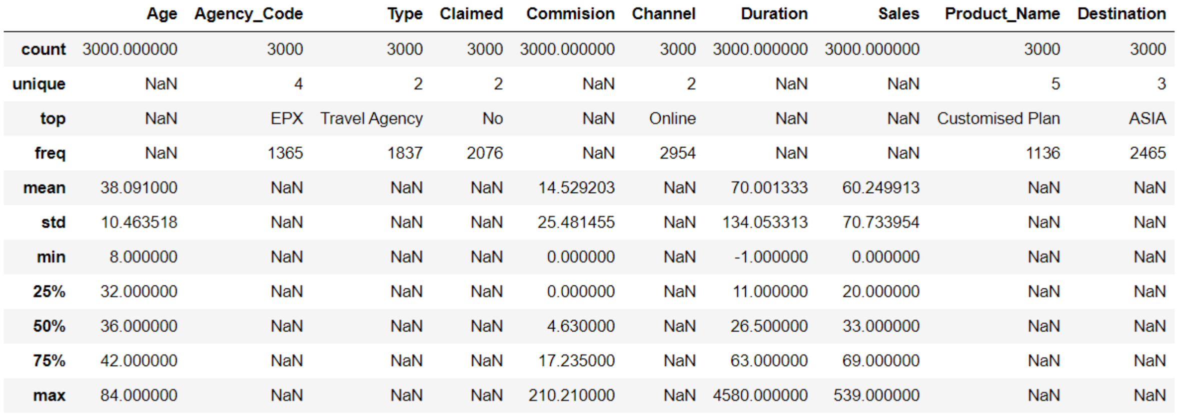
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product_Name	Destination
0	48	C2B	Airlines	No	0.70	Online	7	2.51	Customised Plan	ASIA
1	36	EPX	Travel Agency	No	0.00	Online	34	20.00	Customised Plan	ASIA
2	39	CWT	Travel Agency	No	5.94	Online	3	9.90	Customised Plan	Americas
3	36	EPX	Travel Agency	No	0.00	Online	4	26.00	Cancellation Plan	ASIA
4	33	JZI	Airlines	No	6.30	Online	53	18.00	Bronze Plan	ASIA
5	45	JZI	Airlines	Yes	15.75	Online	8	45.00	Bronze Plan	ASIA
6	61	CWT	Travel Agency	No	35.64	Online	30	59.40	Customised Plan	Americas
7	36	EPX	Travel Agency	No	0.00	Online	16	80.00	Cancellation Plan	ASIA
8	36	EPX	Travel Agency	No	0.00	Online	19	14.00	Cancellation Plan	ASIA
9	36	EPX	Travel Agency	No	0.00	Online	42	43.00	Cancellation Plan	ASIA

The data contains of 3000 rows and 10 columns. The names of the columns are 'Age', 'Agency_Code', 'Type', 'Claimed', 'Commision', 'Channel'.

The below are the data types of the data give:

```
0   Age           3000 non-null   int64
1   Agency_Code   3000 non-null   object
2   Type          3000 non-null   object
3   Claimed       3000 non-null   object
4   Commision     3000 non-null   float64
5   Channel       3000 non-null   object
6   Duration      3000 non-null   int64
7   Sales         3000 non-null   float64
8   Product_Name  3000 non-null   object
9   Destination   3000 non-null   object
```

Below are the descriptive statistics of the given data:



	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product_Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	2954	NaN	NaN	1136	2465
mean	38.091000	NaN	NaN	NaN	14.529203	NaN	70.001333	60.249913	NaN	NaN
std	10.463518	NaN	NaN	NaN	25.481455	NaN	134.053313	70.733954	NaN	NaN
min	8.000000	NaN	NaN	NaN	0.000000	NaN	-1.000000	0.000000	NaN	NaN
25%	32.000000	NaN	NaN	NaN	0.000000	NaN	11.000000	20.000000	NaN	NaN
50%	36.000000	NaN	NaN	NaN	4.630000	NaN	26.500000	33.000000	NaN	NaN
75%	42.000000	NaN	NaN	NaN	17.235000	NaN	63.000000	69.000000	NaN	NaN
max	84.000000	NaN	NaN	NaN	210.210000	NaN	4580.000000	539.000000	NaN	NaN

From the table above we can see that the average age of the people who are present in this data is 38 years. The highest age of a person is 84 years and the smallest person in the data is 8 years old. If we look at the mean of the data which is the average it is 38 years with a standard deviation of 10.

Now when it comes to the duration of the tour, we clearly have an outlier here in this data. We can see that the average mean of the during of the tour is 70 with highest value being 4580 and the lowest with -1. The median for this column is 26.5. As we know that the median is more resistant to outliers we can see that the median of the data is low compared to the mean.

The same problem is happening with the column "Sales", we see that there are outlier in this column as well. The mean of the data is given as 60 but the median is 33 with highest value being 69 and the lowest value being 0. The standard deviation for the data given is 70 which is very high for a mean of 60 which clearly states the discrepancies of outliers.

If we look into the column "Commission" we can see that the minimum value is 0 and the maximum value is 210. However if we look at the median it is just 4.63 with 1st quartile 0 and 3rd quartile 17. The mean of this data is 14.52 with a standard deviation of 25.

Below is the variance calculated:

a)Age - 109.485214

b)Commision - 649.304524

c)Duration - 17970.290762

d)Sales - 5003.292182

Variance measures how far each number in the set is from the mean and thus from every other number in the set.

There are no Null values present in the data.

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

I have used the sklearn model selection and done the splitting of data into train and test set. The test size will be 30% and the train size would be 70%.

The dimensions of the data after the splitting of training and test data are

The train data has 2002 rows and 9 columns and the test data has 859 rows and 9 columns.

I have built a model for CART, Random Forest and also for the Artificial Neural Networks. The coding part can be found from the Jupyter Notebook. I shall mention the final conclusion results for each model separately below:

CART MODEL CONCLUSION:

Train Data:

AUC: 81%

Accuracy: 76%

Precision: 65%

f1-Score: 61%

Test Data:

AUC: 79%

Accuracy: 78%

Precision: 68%

f1-Score: 63%

It has been derived from the CART Model that the Agency_Code is the most important variable for predicting Claim Status.

RANDOM FOREST MODEL CONCLUSION:

Train Data:

AUC: 83%

Accuracy: 78%

Precision: 69%

f1-Score: 63%

Test Data:

AUC: 80%

Accuracy: 78%

Precision: 68%

f1-Score: 62%

Training and Test set results are almost similar, the model is a decent model. Agency_Code is again the most important variable for predicting Chaim Status.

ARTIFICIAL NEURAL NETWORKS MODEL CONCLUSION:

Train Data:

AUC: 76%

Accuracy: 75%

Precision: 75%

f1-Score: 45%

Test Data:

AUC: 75%

Accuracy: 73%

Precision: 69%

f1-Score: 39%

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

Below is the performance metrics for individual models

CART MODEL:

Confusion Matrix for the training data using CART Model:

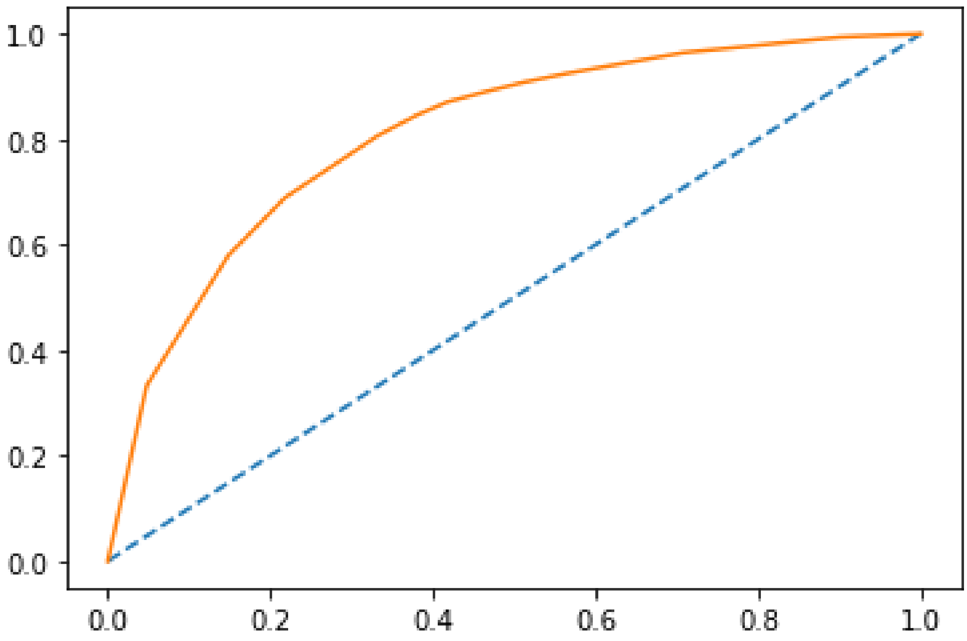
```
array([[1157, 202],
       [ 270, 373]]
```

Confusion Matrix for the test data using CART Model:

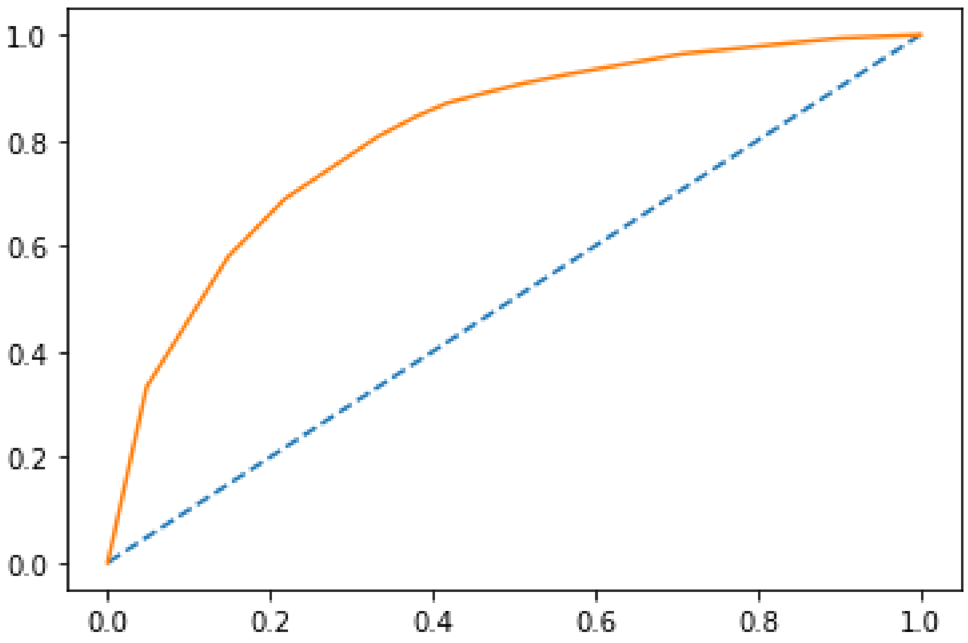
```
array([[510,78],
       [ 109,162]]
```

The accuracy for the train data is 76% and for the test data is 78%.

Below is the ROC Curve for CART Model for training data:



Below is the ROC Curve for CART Model for testing data:



The AUC Score for training data is: 81%

The AUC Score for testing data is: 79%

RANDOM FOREST MODEL:

Confusion Matrix for the training data using CART Model:

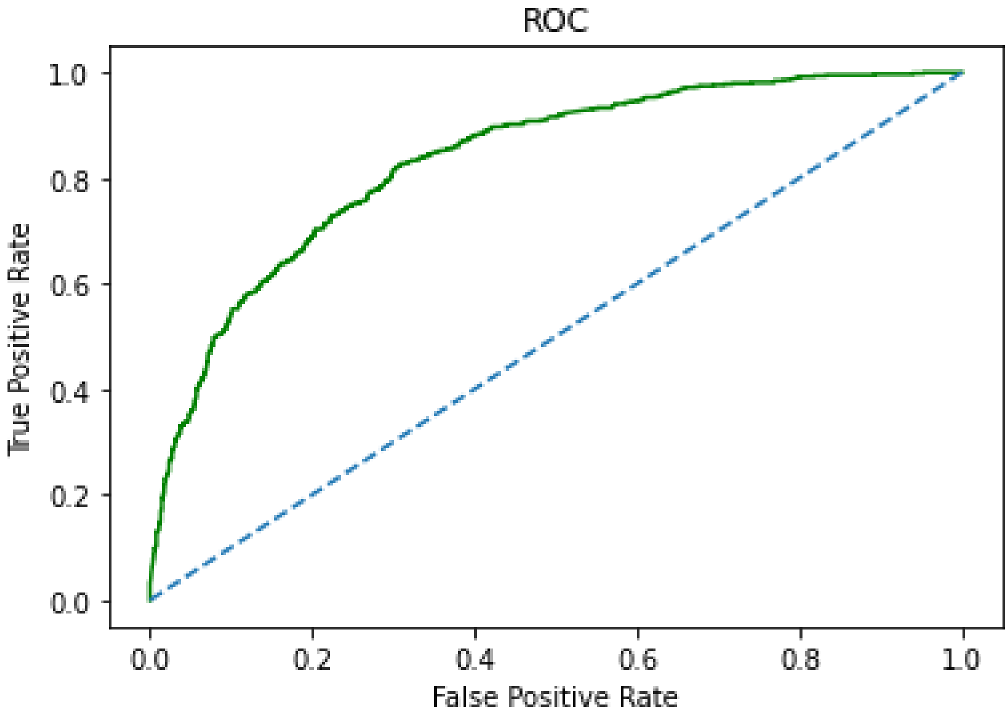
```
array([[1191, 168],
       [ 269, 374]]
```

Confusion Matrix for the test data using CART Model:

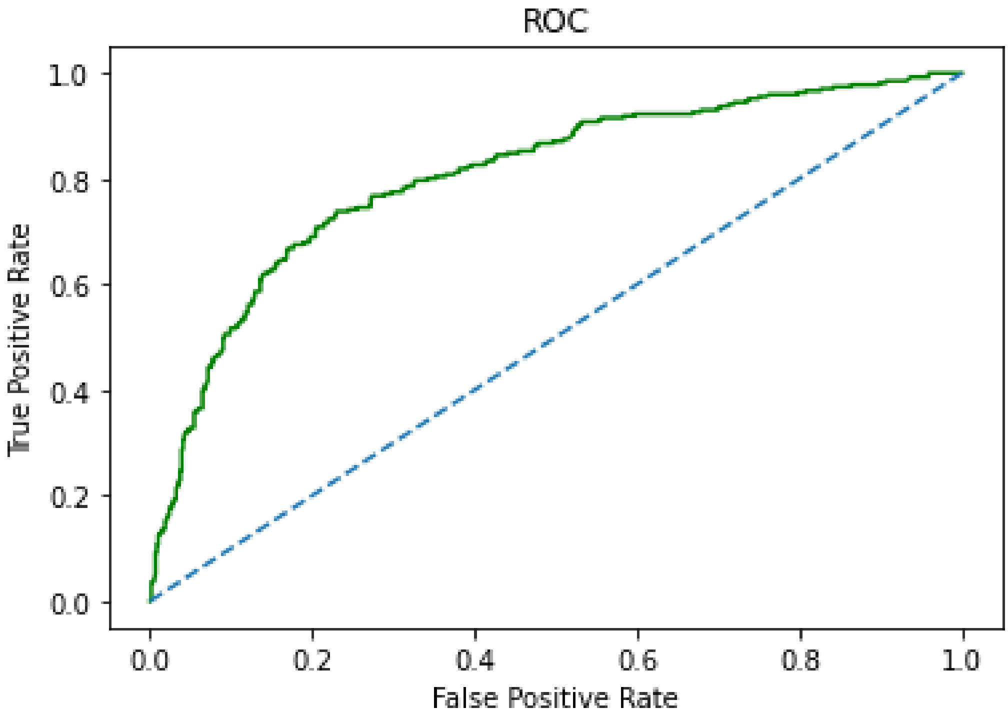
```
array([[513,75],
       [ 115,156]]
```

The accuracy for the train data is 78% and for the test data is 78%.

Below is the ROC Curve for Random Forest Model training data:



Below is the ROC Curve for Random Forest Model testing data:



The AUC Score for training data is: 83%

The AUC Score for testing data is: 80%

ARTIFICIAL NEURAL NETWORKS MODEL:

Confusion Matrix for the training data using ANN Model:

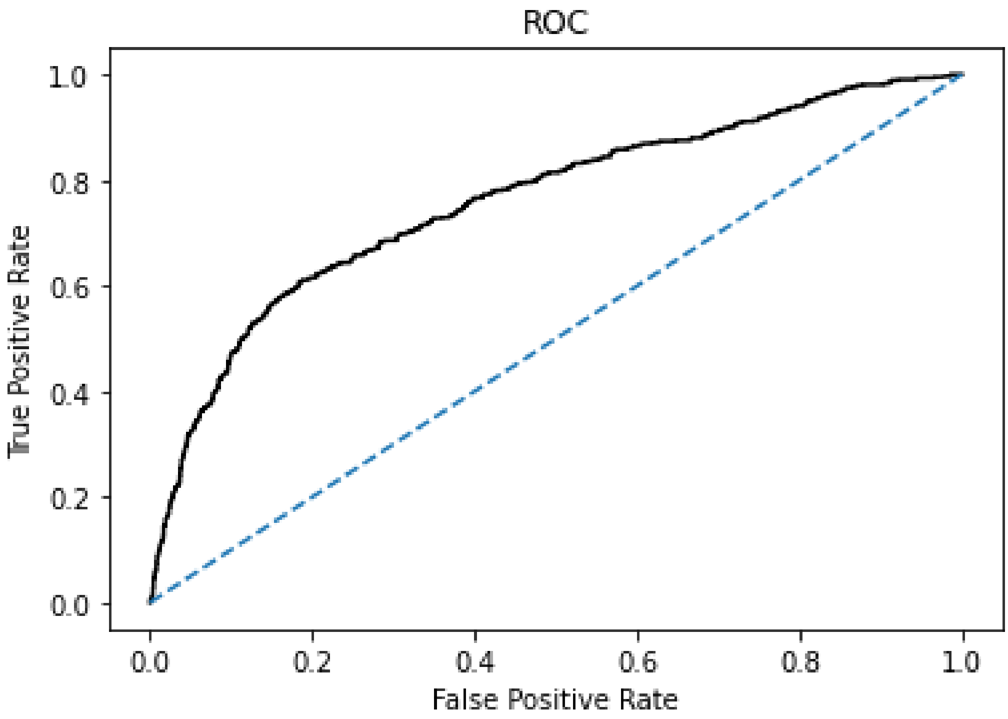
```
array([[1292, 67],
       [ 437, 206]]
```

Confusion Matrix for the test data using ANN Model:

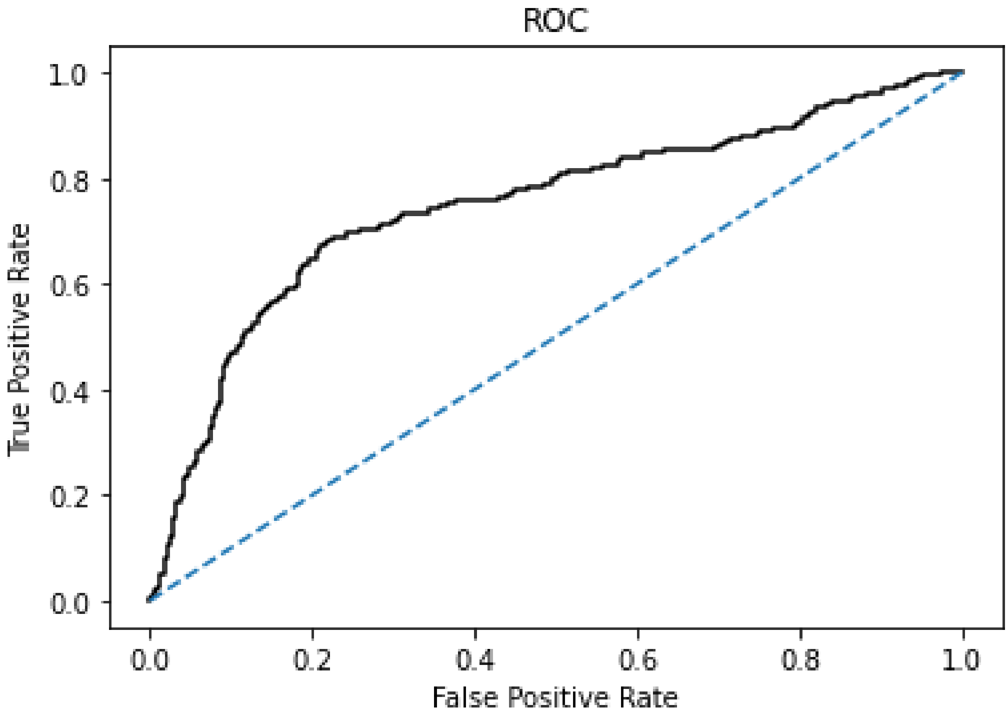
```
array([[554,34],
       [ 197,74]]
```

The accuracy for the train data is 75% and for the test data is 73%.

Below is the ROC Curve for Random Forest Model training data



Below is the ROC Curve for Random Forest Model testing data:

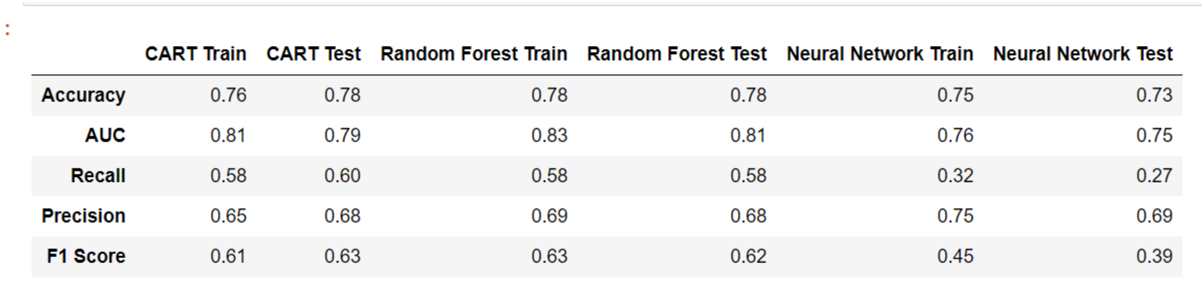


The AUC Score for training data is: 76%

The AUC Score for testing data is: 75%

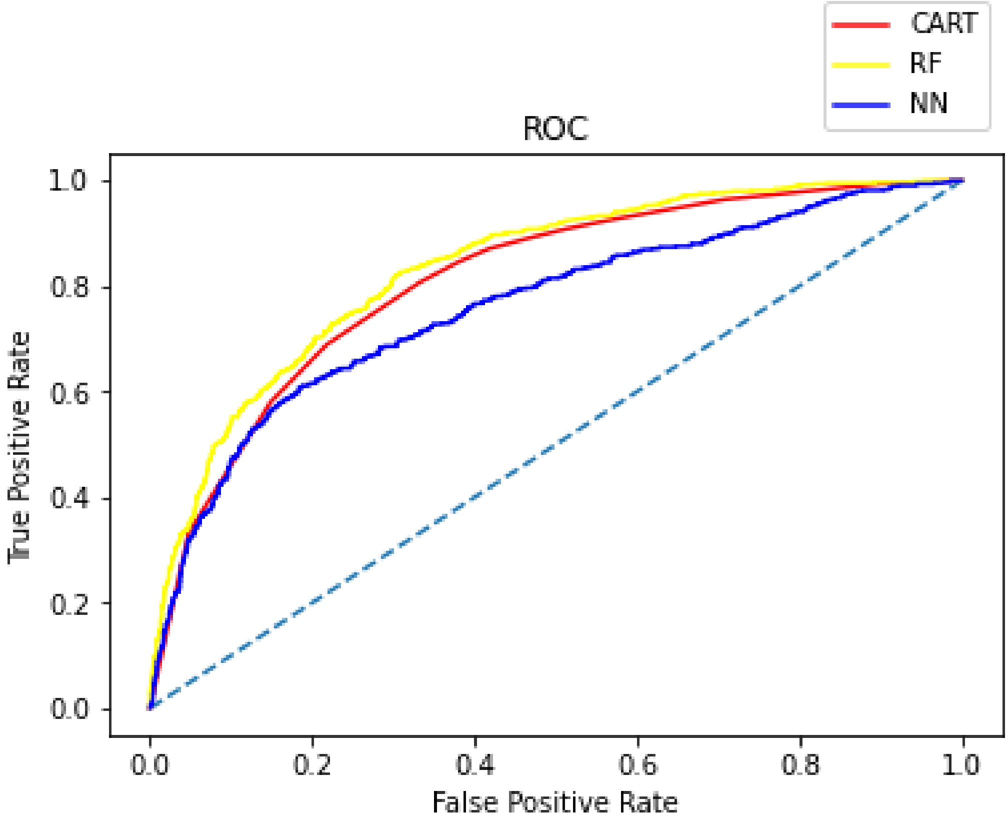
2.4 Final Model: Compare all the model and write an inference which model is best/optimized.

Comparison of the performance metrics from the 3 models:

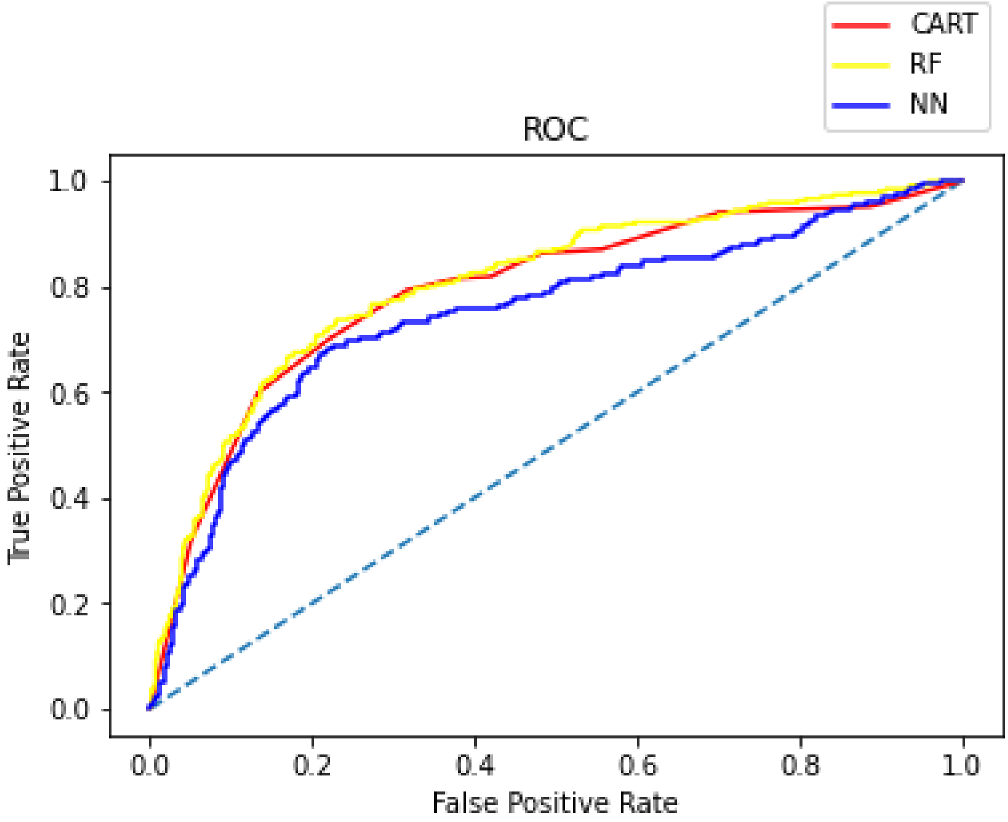


	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.76	0.78	0.78	0.78	0.75	0.73
AUC	0.81	0.79	0.83	0.81	0.76	0.75
Recall	0.58	0.60	0.58	0.58	0.32	0.27
Precision	0.65	0.68	0.69	0.68	0.75	0.69
F1 Score	0.61	0.63	0.63	0.62	0.45	0.39

ROC Curve for the 3 models on the Training data:



ROC Curve for the 3 models on the Test data:



Out of the 3 models, Random Forest has slightly better performance than the Cart and Neural network model

Overall all the 3 models are reasonably stable enough to be used for making any future predictions. From Cart and Random Forest Model, the variable Agency_Name is found to be the most useful feature amongst all other features for predicting if a person will claim the insurance or not.

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

From the whole analysis, I’ve come to a conclusion where we can predict that the Agency code is one of the important factor in predicting if the person will claim the insurance or not.

There are four different types of agency code : 'C2B', 'EPX', 'CWT', 'JZI'

If the insurance company can find out which Agency code is having the maximum amount of insurance claims and maybe they can look into that tour firm with special interest and then make changes which would be beneficial for the insurance firm. Then maybe they can understand where exactly the problem is and maybe they can rectify it.