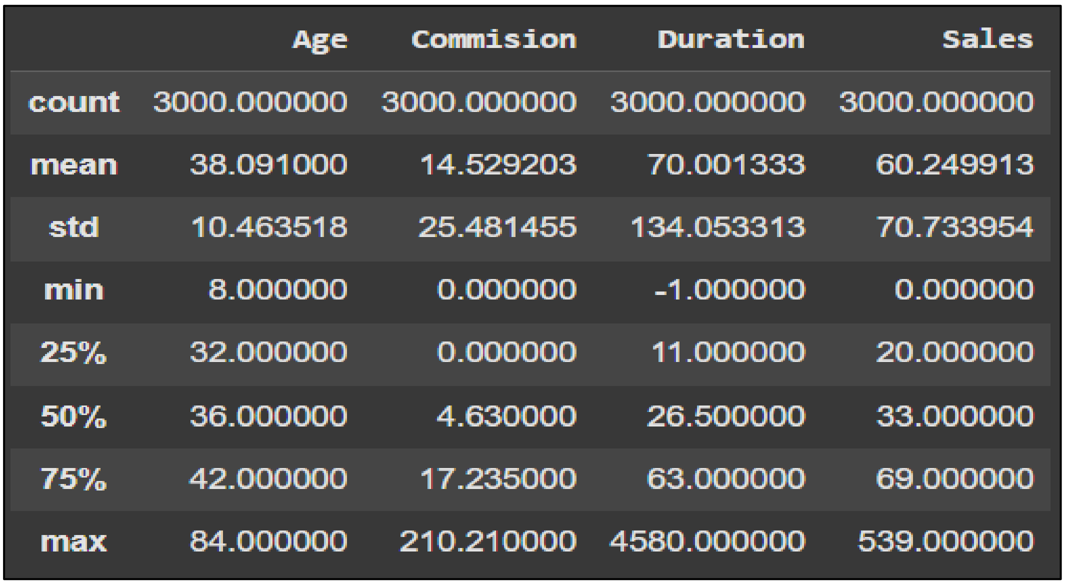
Final Report

By: Parikshith M (FT253060)

Objective: To predict the “Claim” status of a tour insurance firm using CART and Random Forest and compare their performances to derive insights and make informed decisions to reduce the claim frequency.

Steps involved in building the model

1. Understanding the Data: Univariate Analysis



	Age	Commision	Duration	Sales
count	3000.000000	3000.000000	3000.000000	3000.000000
mean	38.091000	14.529203	70.001333	60.249913
std	10.463518	25.481455	134.053313	70.733954
min	8.000000	0.000000	-1.000000	0.000000
25%	32.000000	0.000000	11.000000	20.000000
50%	36.000000	4.630000	26.500000	33.000000
75%	42.000000	17.235000	63.000000	69.000000
max	84.000000	210.210000	4580.000000	539.000000

- **AGE:** For the Age variable: IQR = Q3- Q1 = 34

  TO see the outlier at an outer extent, we can use Q1-1.5IQR and Q3 + 1.5 IQR, which gives us a values of 20 and 54, but 20 and 54, age seems to be okay, and we do not consider this as an outlier.

  we can see that the median age is 36, so we can deduce that people in the mid-30s are the probable age grouped people who tend to buy insurance

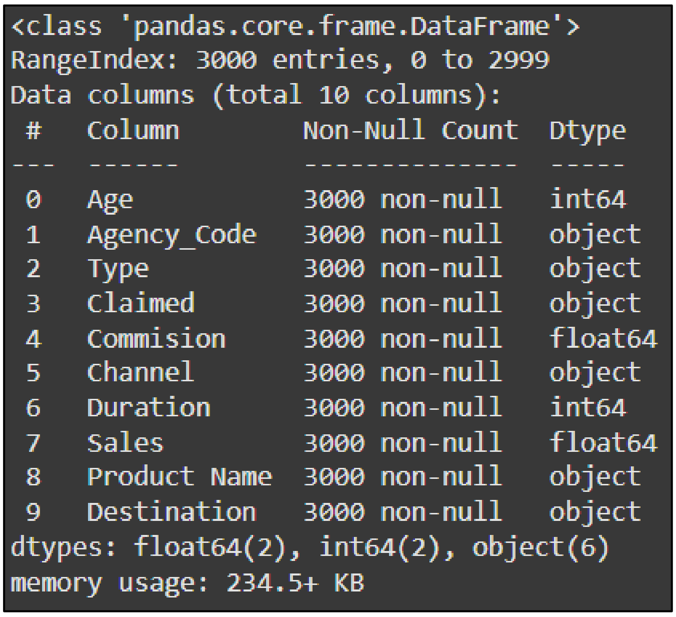
- **Commission**: For the commission column, we can see the min and 25th percentile values to be 0, this might look like a null value, but, in general we can have 0 commission, hence we can keep the column as it is, without any further treatment.
- **Duration:** For the Duration column, we have min value as -1, which seems an error in the data, we need to identify such records and remove them, also maximum value of 4580 seems illogical and impossible to achieve.

So, we can remove that particular row as well from the data set for our further analysis.

We can also see that the median duration is 26 days, from this we can deduce the trip to last around a month on an average, we can also use this data to identify the time during which people usually go on vacations to classify better.

- **Sales:** For the sales column, we see the min value of 0, but in general, if the sales value is 0, it doesn't make sense when a travel insurance is claimed in such a case, so we can proceed to delete the rows that has sales values as 0.

  we can see for all the values, mean is greater than median, that indicate that the data in all the columns are right skewed.



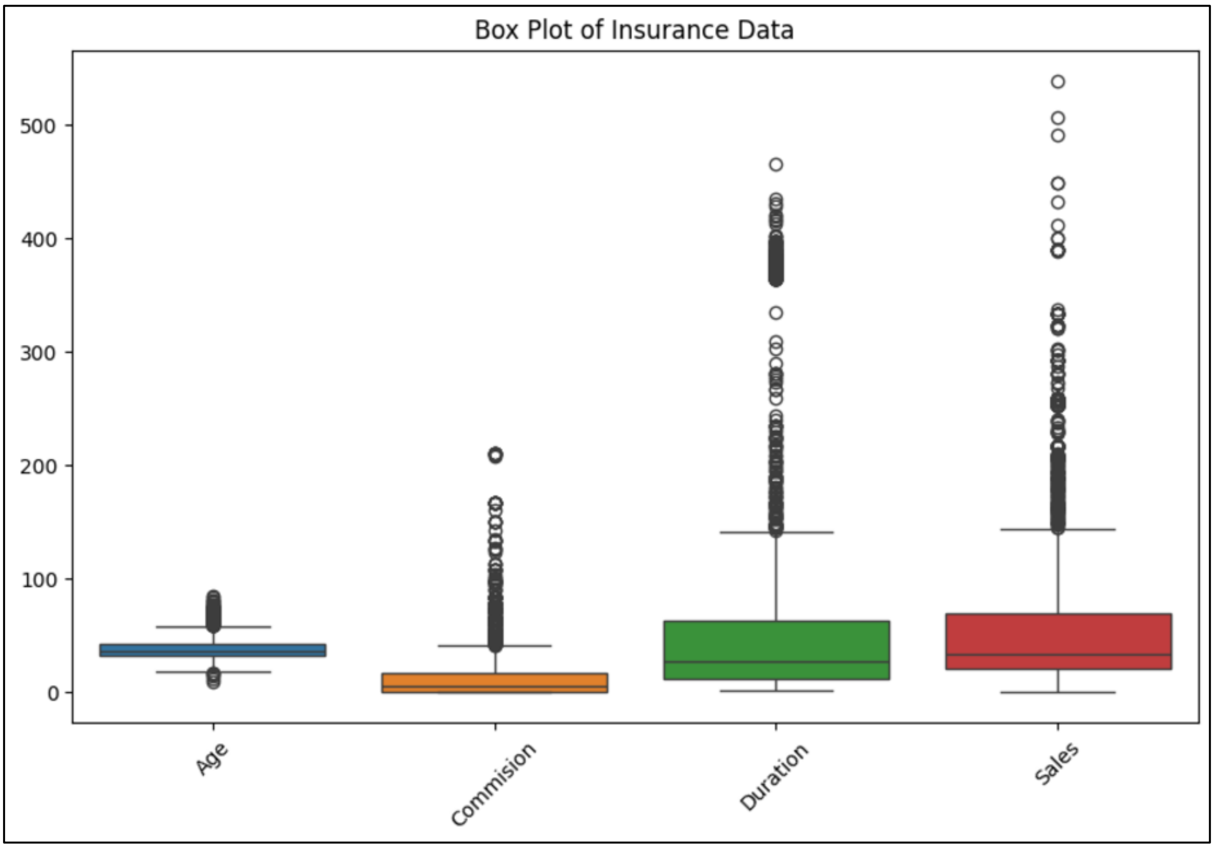
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 10 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Age           3000 non-null   int64  
 1   Agency_Code   3000 non-null   object 
 2   Type          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 
dtypes: float64(2), int64(2), object(6)
memory usage: 234.5+ KB
```

- In total there are 3000 rows and 10 columns in the data, this data set is using a memory of 234.5 KB. All the columns look good with respect to the data type that it has been associated with, there are no mismatches with respect to the column and the data type of that column

Boxplot:

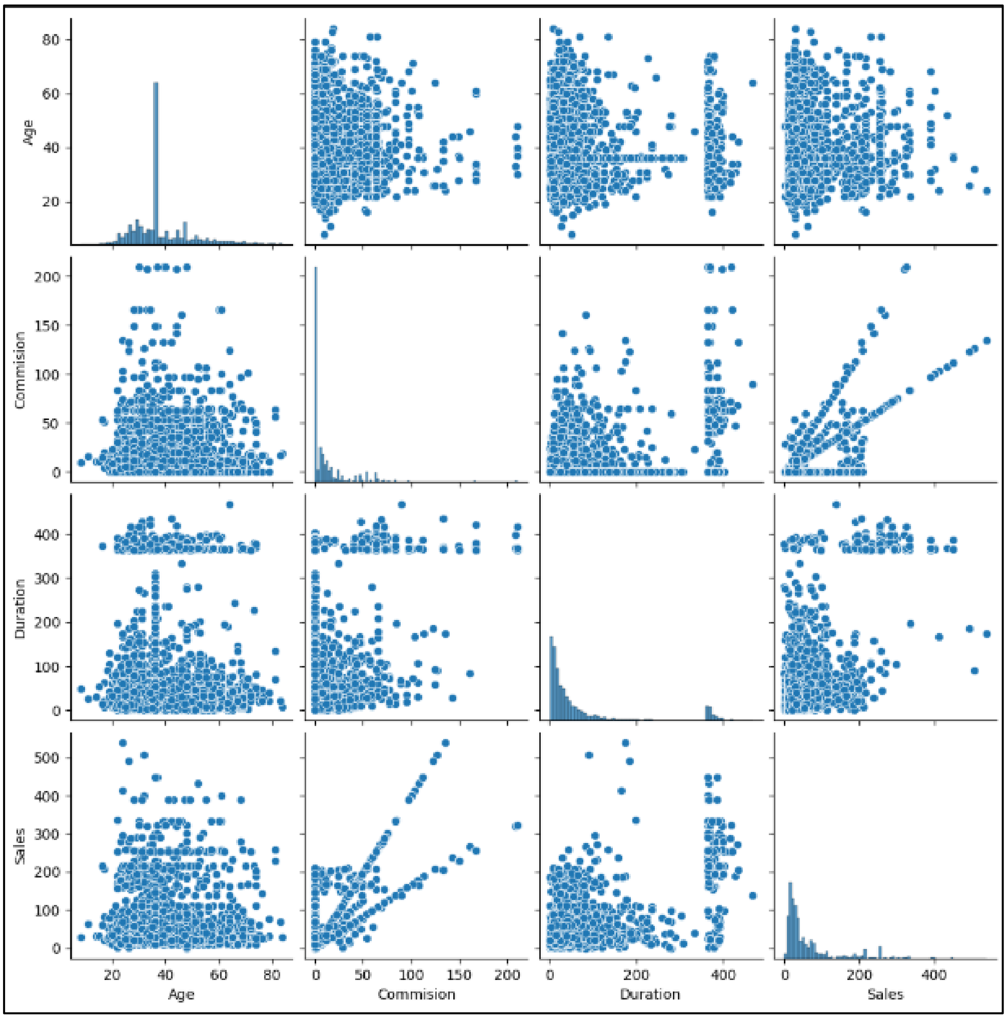
- It can be observed from the Box plot that all the numerical columns are right skewed.
- Median value of commission is almost near 0, indicating not every sale leads to a commission.
- The duration has a median of around 30, which is also what we can observe from the initial observations.

- In all the variables considered, it can be seen that there are many values that are greater than the 75th percentile value, i.e Q3, even though they seem like outliers, they are valid values and we need not treat them.

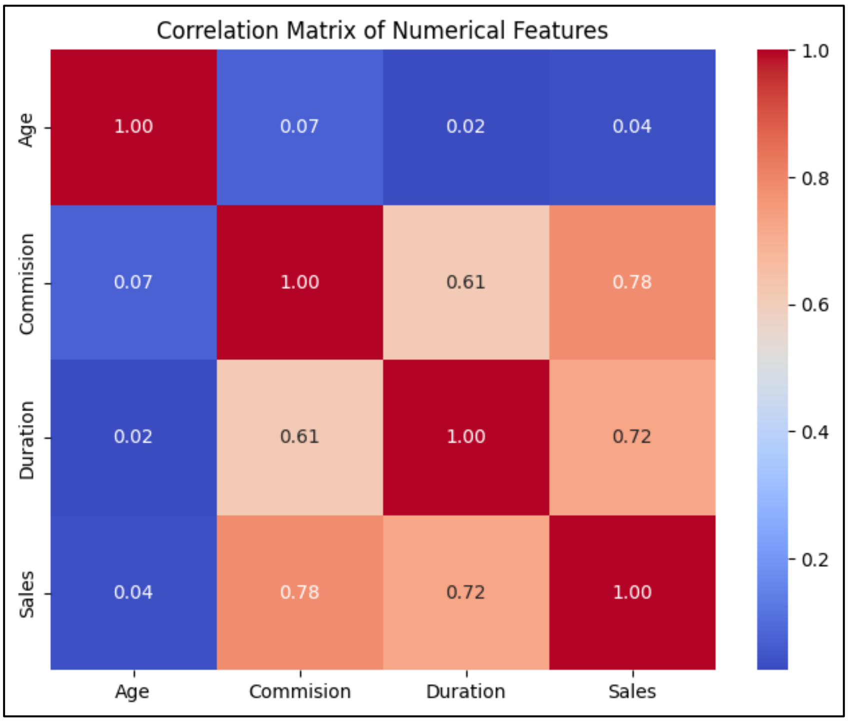


Bi-Variate and Multi-Variate Analysis:

AGE*:* From the Pair plot we can see that, Age doesn't show any visible correlation with any other variable. The distribution is more haphazard and doesn't show any trend.
Commission*:* Here we can see that commission shows a behaviour which seems to be correlated with sales, which also seems logical, as commission is a variable expressed in terms of percentage of sales
Duration*:* Here duration doesn't seem to be correlated with any other variable, the plot has a cloud like structure, so we don't see any correlation. Here, we can observe a presence of multiple classes, which indicates a cluster of short-trips and long-trips


Age
Commision
Duration
Sales
80
60
40
20
200
150
100
50
0
400
300
200
100
0
500
400
300
200
100
0
20
40
60
80
Age
0
50
100
150
200
Commision
0
200
400
Duration
0
200
400
Sales

Heatmap:


Correlation Matrix of Numerical Features
Age
Commision
Duration
Sales
1.00
0.07
0.02
0.04
0.07
1.00
0.61
0.78
0.02
0.61
1.00
0.72
0.04
0.78
0.72
1.00
Age
Commision
Duration
Sales
1.0
0.8
0.6
0.4
0.2

- High Correlation value of 0.72 can be observed between duration and sales, which indicates duration and sales are linearly correlated with each other, indicating higher duration leads to higher sales value
- Sales and commission also have a high correlation value of 0.78, as commission is expressed as a percentage of sales, this behaviour is expected.

2. Data Treatment:

- Duration column having values -1,0 and 4580 are removed. Sales column with value 0, which indicates that the sale has not happened, so claiming insurance for a sale that has not happened doesn’t make sense.
- If the sales are 0, then it might indicate that the person might have cancelled the Plan before actual Delivery or they might have been offered with complementary deals. So, 53 such rows were identified and removed.
- Size of the data after data treatment,
  The number of rows (observations) is: **2943**
  The number of columns(variables) is: **10**

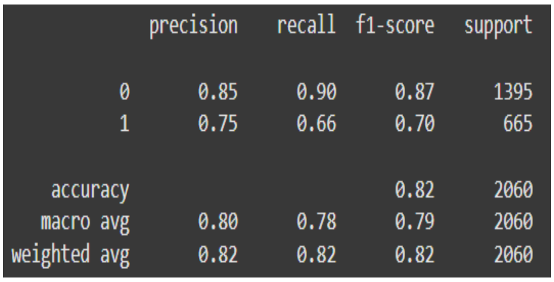
3. Building model using CART:

- The data is split into 70% train and 30% test data.
- One hot Encoding method ss used to convert the Categorical variables to numerical format.
- **Claimed** is identified to be the Y (Target) variable.
- Calculated the Gini score values to identify the best features based on its values.
- The Model had an accuracy of 99% on the train data and 71.5% on the test data, which clearly indicates that the model is completely overfitting.
- Pruning is used to rectify the issue of overfitting.

Evaluating the model after Pruning:

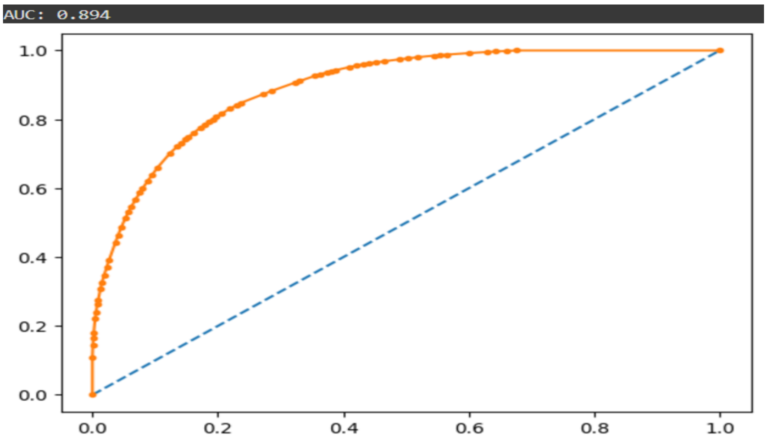
On Train Data: This model has an accuracy of 82%, this has solved the problem of Overfitting.

It has a very good specificity value of 90% and subpar sensitivity value of 66



```
              precision    recall  f1-score   support

           0       0.85      0.90      0.87      1395
           1       0.75      0.66      0.70       665

    accuracy                           0.82      2060
   macro avg       0.80      0.78      0.79      2060
weighted avg       0.82      0.82      0.82      2060
```

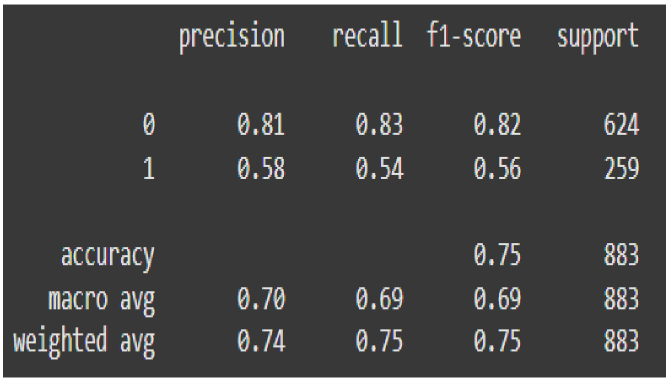
ROC CURVE FOR THE TRAIN DATA:



```
AUC: 0.894
```

The AUC curve shows a value of **89%**, which indicates that the model has a probability of **89%** in predicting the correct class for the positive and negative class values.

Performance on the Test Data:

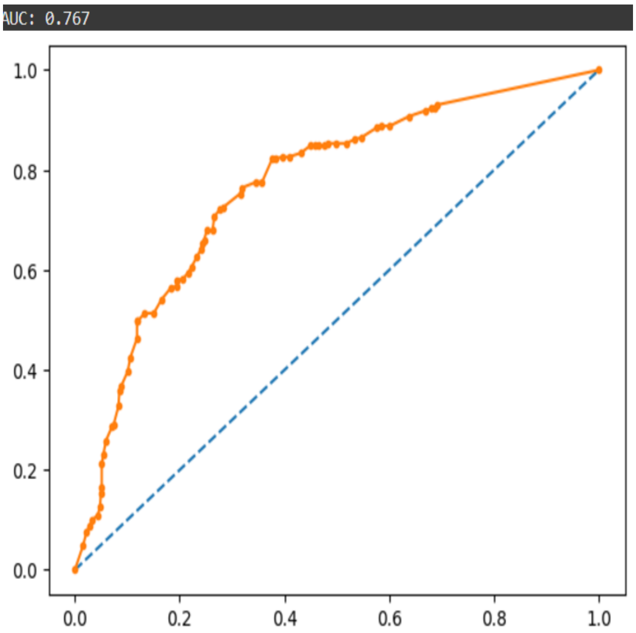


```
              precision    recall  f1-score   support

           0       0.81      0.83      0.82       624
           1       0.58      0.54      0.56       259

    accuracy                           0.75       883
   macro avg       0.70      0.69      0.69       883
weighted avg       0.74      0.75      0.75       883
```

The model has an accuracy of **75%**, but performs better for the negative claim class of 0.

There is a significant drop in the prediction of the class 1 claimed class for the test data.

ROC Curve for the test data:



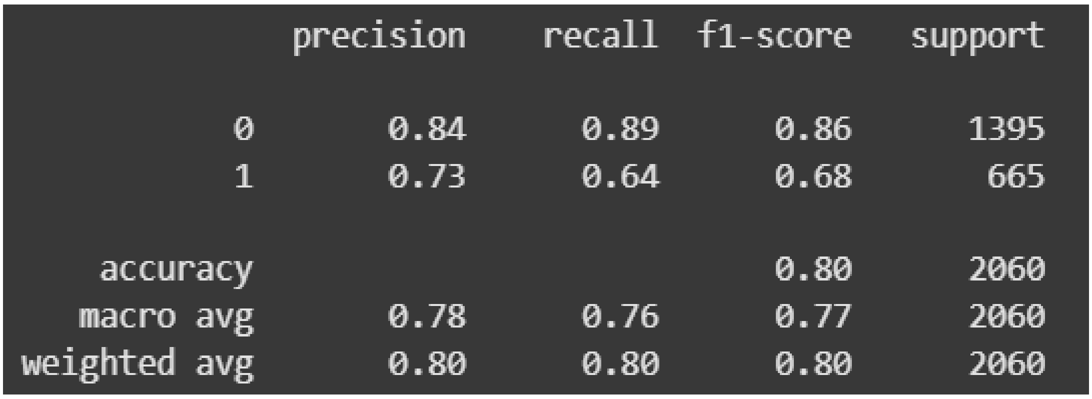
```
AUC: 0.767
```

THE AUC curve shows a value of **77%**, which indicates that the model has a probability of **77%** in predicting the correct class for the positive and negative class values. Hence this is a decent model in predicting the classes, and predicting the classes for the Claimed variable**.**

4. Building Model using Random Forest:

- The data is split into 70% train and 30% test data.
- One hot Encoding method is used to convert the Categorical variables to numerical format.
- **Claimed** is identified to be the Y (Target) variable.
- Calculated the Gini score values to identify the best features based on its values.
- The Model had an accuracy of 99.7% on the train data and 73% on the test data, which clearly indicates that the model is completely overfitting.
- To address the issue of overfitting, Grid Search Method is used.
- Grid Parameters set:
  param_grid = {
  max_depth': [5,10,15],
  max_features': [2,4],
  min_samples_leaf': [10,100],
  min_samples_split': [10,50,200],
  n_estimators': [101,301,501]
- **Evaluating the model after grid search:**

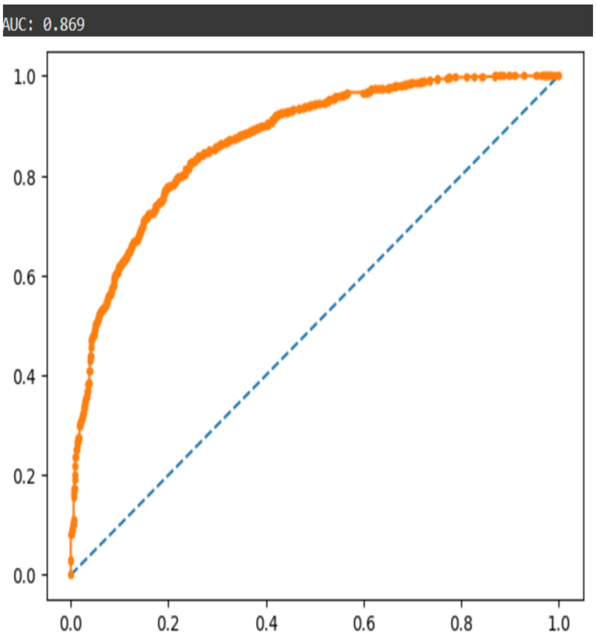
Evaluation on the train data:



```
              precision    recall  f1-score   support

           0       0.84      0.89      0.86      1395
           1       0.73      0.64      0.68       665

    accuracy                           0.80      2060
   macro avg       0.78      0.76      0.77      2060
weighted avg       0.80      0.80      0.80      2060
```

- The Train Data gave out an Accuracy of **80%**, indicating that the issue of overfitting is now resolved with the help of the Grid search that we used to identify the best parameters.
- **80%** is an above par value and this seems like a good model for the Train data.
- The model has a specificity of **89%**, which is excellent in predicting the Class 0 **claimed** values, but it has a sub-par sensitivity value of 64%, in identifying the class 1 **claimed** values.

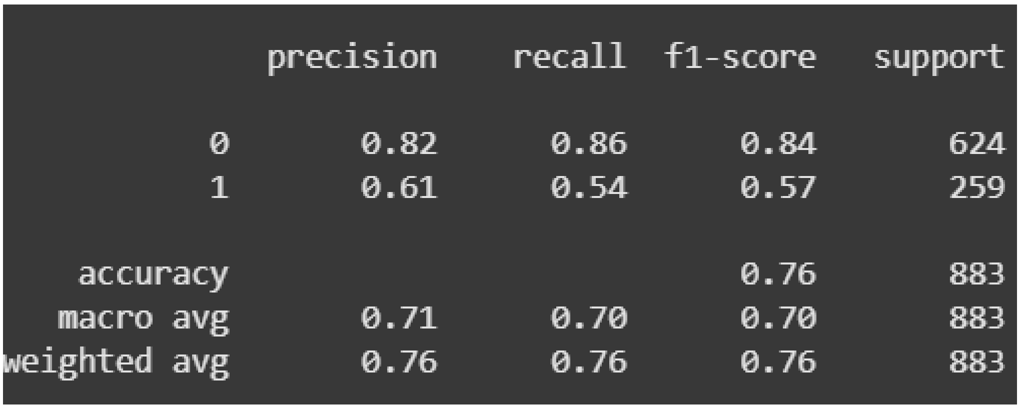
Roc Curve for the Train Data:



```
AUC: 0.869
```

From the ROC cure we see that the value we have got is **86.9**%, which indicates that the model has a **86.9**% chance of correctly distinguishing between the positive class (Claimed = 1) and the negative class (Claimed =0), so this is an indicator of a good model.

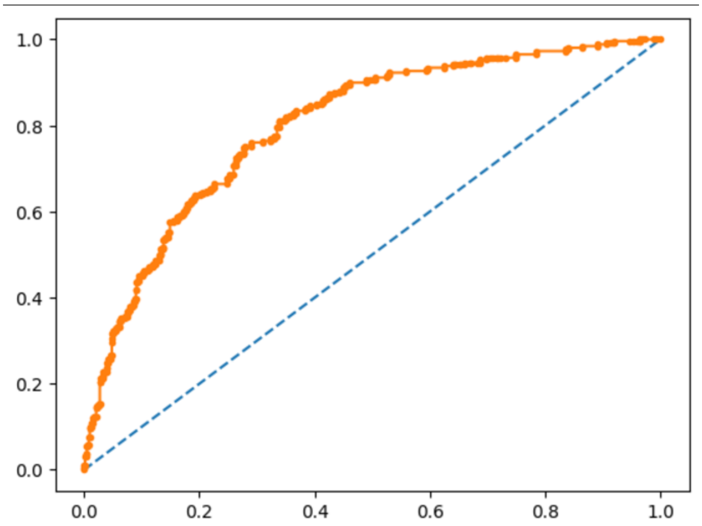
Evaluation on the Test Data:



```
              precision    recall  f1-score   support

           0       0.82      0.86      0.84       624
           1       0.61      0.54      0.57       259

    accuracy                           0.76       883
   macro avg       0.71      0.70      0.70       883
weighted avg       0.76      0.76      0.76       883
```

- The Model has an accuracy of 76% on the Test Data, which is an Above Par Model.
- The Model is doing a really job in predicting the Class 0 problems i.e (Claimed=0) which is indicated by the Specificity of 86%, but does a poor job on the Class 1 problem i.e (Claimed =1) which is indicated by a sensitivity of 54%.

ROC Curve for the Test Data:



Here we can observe that the, AUC value is 80%, which indicates that the model has a 80% chance of correctly distinguishing between the positive class (Claimed = 1) and the negative class (Claimed =0), so this is an indicator of a good model

<u>Recommendations :</u>

Model Choice:

- Random Forest (RF) is the model that I would be recommending here. This model performs better when compared to the decision tree model especially for the test data.
- Random Forest also has a better ROC-AUC value of 80% compared to the 77% value of CART.
- The only improvement with this model is with respect to the sensitivity, other than that the RF model looks balanced and usable for our purpose of prediction.

Business Insights:

- More focus and emphasis can be given to the class of people aged 30-40, and those with tours that has a higher duration as these people are more prone to claims.
- As there is a strong correlation between sales and duration, so tours that have higher duration can be considered as higher risk. Insurance premium can be revised for these.
- Insurance agents need to prioritise risk assessment rather than solely focusing on commission.