Problem 2 Statement

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.

Exploratory Data Analysis :

Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
48	C2B	Airlines	No	0.7	Online	7	2.51	Customised Plan	ASIA
36	EPX	Travel Agency	No	0	Online	34	20	Customised Plan	ASIA
39	CWT	Travel Agency	No	5.94	Online	3	9.9	Customised Plan	Americas
36	EPX	Travel Agency	No	0	Online	4	26	Cancellation Plan	ASIA
33	JZI	Airlines	No	6.3	Online	53	18	Bronze Plan	ASIA
45	JZI	Airlines	Yes	15.75	Online	8	45	Bronze Plan	ASIA
61	CWT	Travel Agency	No	35.64	Online	30	59.4	Customised Plan	Americas
36	EPX	Travel Agency	No	0	Online	16	80	Cancellation Plan	ASIA
36	EPX	Travel Agency	No	0	Online	19	14	Cancellation Plan	ASIA
36	EPX	Travel Agency	No	0	Online	42	43	Cancellation Plan	ASIA
37	C2B	Airlines	Yes	46.96	Online	368	187.85	Silver Plan	ASIA
43	C2B	Airlines	Yes	15.88	Online	77	63.5	Silver Plan	ASIA
36	EPX	Travel Agency	No	0	Online	23	110	Customised Plan	EUROPE
52	C2B	Airlines	Yes	5.88	Online	7	23.5	Bronze Plan	ASIA
31	CWT	Travel Agency	No	23.76	Online	21	39.6	Customised Plan	ASIA
39	C2B	Airlines	Yes	54	Online	366	216	Silver Plan	ASIA
36	EPX	Travel Agency	No	0	Online	2	42	Customised Plan	ASIA
45	CWT	Travel Agency	No	59.4	Online	40	99	Customised Plan	Americas
23	JZI	Airlines	No	18.2	Online	33	52	Bronze Plan	ASIA

We are provided with the above data set of 3000 rows and 10 columns. Among the given variables , 6 are object type data, 2 are integer type data and 2 are float type data.

The data has no Null values.

```
<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
```

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

From the descriptive stats of the data set,

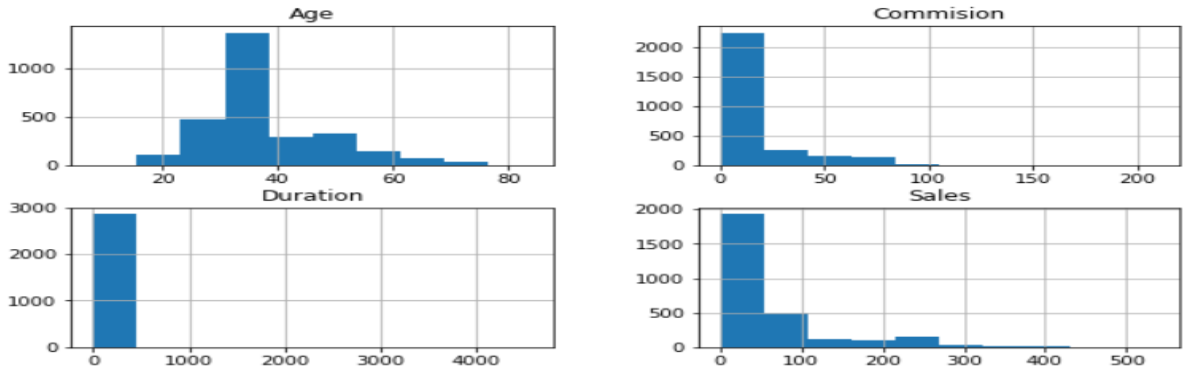
	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000.000000	3000	3000	3000	3000.000000	3000	3000.000000	3000.000000	3000	3000
unique	NaN	4	2	2	NaN	2	NaN	NaN	5	3
top	NaN	EPX	Travel Agency	No	NaN	Online	NaN	NaN	Customised Plan	ASIA
freq	NaN	1365	1837	2076	NaN	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

All the variables have a **value count** of **3000** with **no null values**.

The data set has 139 duplicate rows which have to be dropped.

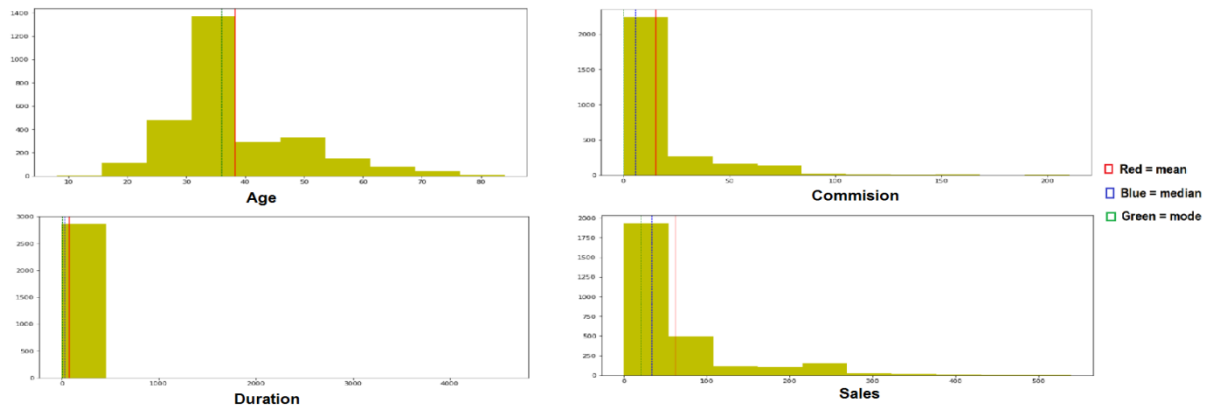
Of the entire dataset, column **'Duration'** has both **highest max** value of 4580 and **least min** value of -1

Columns **'Duration'** and **'Commision'** have highest mean value – 70.00 and least mean value –14.529 respectively.

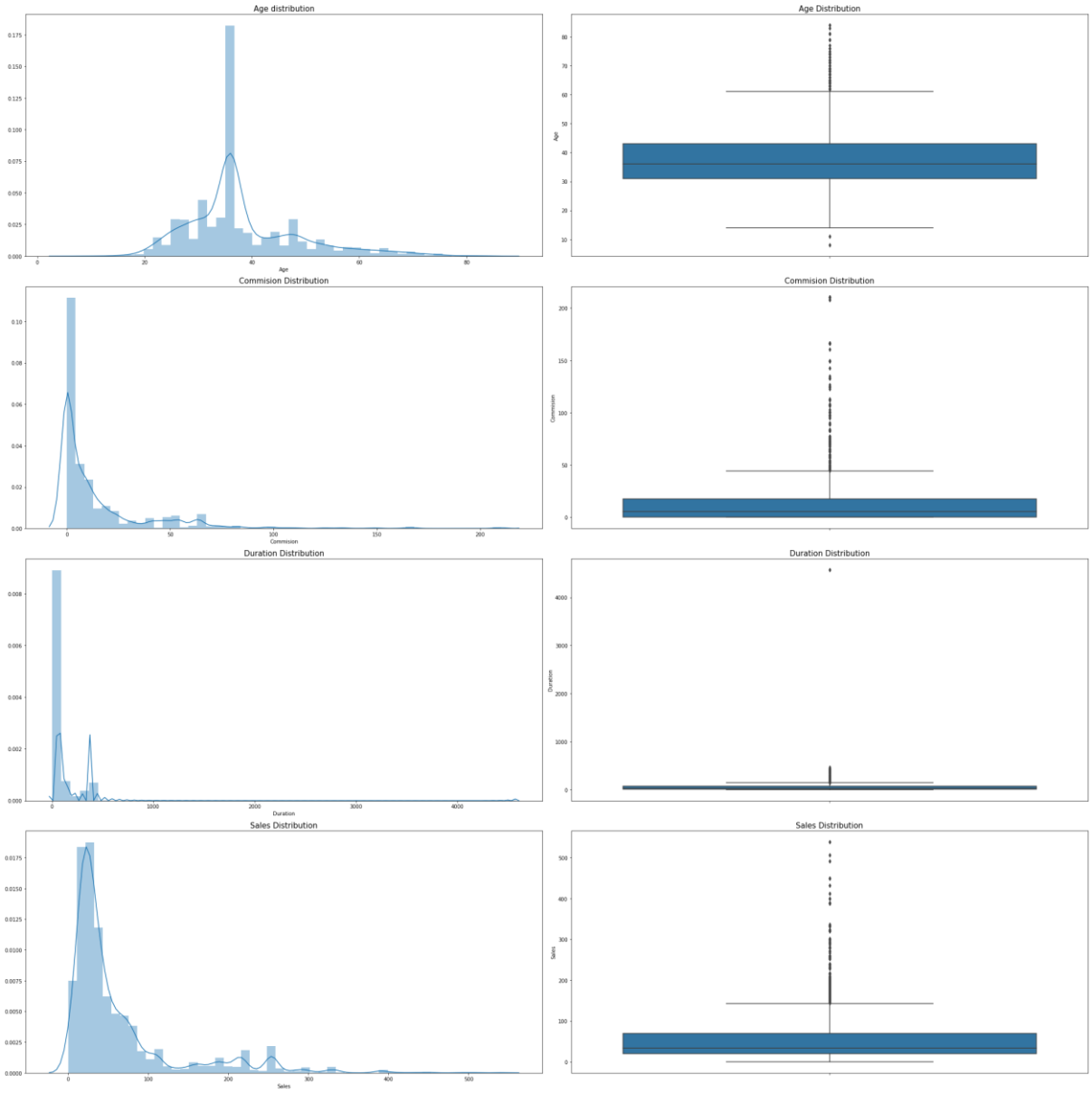


From the above histograms of the variables , we can see that all of the variables are not symmetrical.

Among all the variables , Variable **'Duration'** is most **right skewed** (skew = 13.786) and Variable **'Age'** is least **right skewed** (skew = 1.103).



All the numerical variables of the given data set are having outliers.



Although outlier treatment is not necessary for ANN, Random Forest as they are not sensitive to outliers. We can go ahead with outlier treatment for simplicity.

Multivariate Analysis:

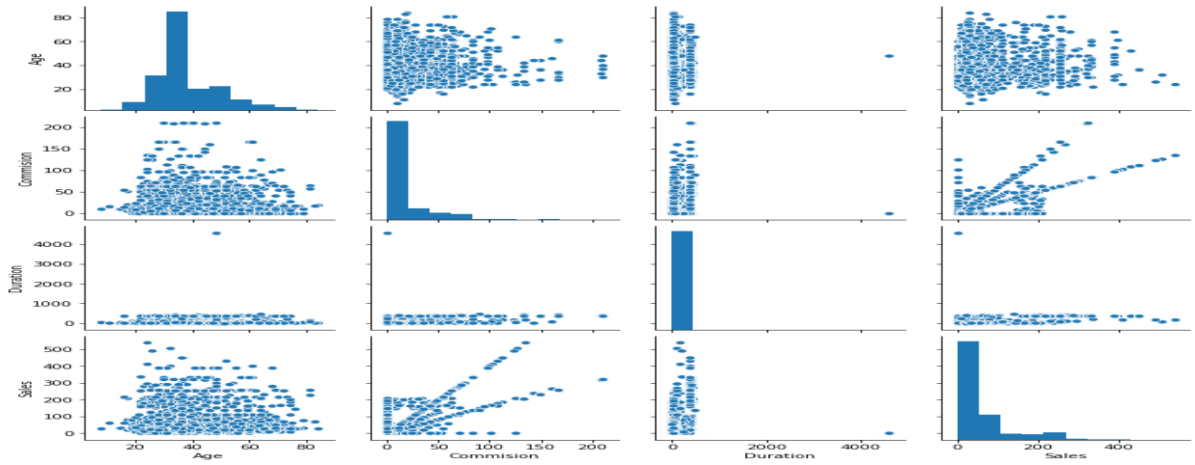
We have the following correlation among the different variables given in the dataset.

	Age	Commision	Duration	Sales
Age	1.000000	0.064759	0.027457	0.036187
Commision	0.064759	1.000000	0.462114	0.762181
Duration	0.027457	0.462114	1.000000	0.549889
Sales	0.036187	0.762181	0.549889	1.000000

HeatMap:

From the above map , we can see that many columns are co-related to each other and there is **highest correlation** (0.76) between variables **'Sales'** and **'Commision'**. Also there is **least correlation**(0.027) between variables **'Duration'** and **'Age'.**

Pairplot:



In the above plot scatter diagrams are plotted for all the numerical columns in the dataset. From the visual representation , we can understand the degree of correlation between any two columns of the given dataset.

Variable **'Sales'** shows high positive linear correlation with Variable **'Commision'**.

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

After dropping duplicate rows, treating outliers and converting all the variables into integer data type.The final data set has the 2861 rows of data.

There is no issue of class imbalance here as we have reasonable proportions in both the classes in the target column **'Claimed'.**

```
0    0.680531
1    0.319469
Name: Claimed, dtype: float64
```

Final dataset details:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2861 entries, 0 to 2999
Data columns (total 10 columns):
 #    Column          Non-Null Count   Dtype
---   ------          --------------   -----
 0    Age             2861 non-null    float64
 1    Agency_Code     2861 non-null    int8
 2    Type            2861 non-null    int8
 3    Claimed         2861 non-null    int8
 4    Commision       2861 non-null    float64
 5    Channel         2861 non-null    int8
 6    Duration        2861 non-null    float64
 7    Sales           2861 non-null    float64
 8    Product Name    2861 non-null    int8
 9    Destination     2861 non-null    int8
dtypes: float64(4), int8(6)
memory usage: 208.5 KB
```

We have split the data in the ratio 70:30 for train and test datasets.

```
X_train (2002, 9)
X_test (859, 9)
train_labels (2002,)
test_labels (859,)
```

Decision Tree Classifier:

This supervised learning method is useful for classification and regression. The generated model will help in predicting the target variable through learning simple decision rules.

We got the following best parameters through GridSearchCV for the dataset,

DecisionTreeClassifier (max_depth=**8**, min_samples_leaf=**30**, min_samples_split=**240**)

The highest importance feature denoted by this method is **Agency_Code (57.79% importance)**

Train Data:
AUC: 82%
Accuracy: 77%
Sensitivity: 52%
Precision: 70%
f1-Score: 60%

Test Data:
AUC: 79%
Accuracy:78%
Sensitivity: 51%
Precision: 71%
f1-Score: 59%

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

Agency_Code is the most important variable for predicting claim status

Random Forest Classifier:
Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

We got the following best parameters through GridSearchCV for the dataset,

RandomForestClassifier(max_depth=**10**, max_features=**8**, min_samples_leaf=**30**, min_samples_split=**90**, n_estimators=**200**)

The highest importance feature denoted by this method is **Agency_Code (44.93% importance).**

Train Data
AUC: 73%
Accuracy: 79%
Sensitivity: 59%
Precision: 69%
f1-Score: 64%

Test Data
AUC: 73%
Accuracy: 78%
Sensitivity: 58%
Precision: 69%
f1-Score: 63%

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

Agency_Code is again the most important variable for predicting claim status.

Neural Network Classifier:

A Neural Network is a computational model loosely based on the functioning cerebral cortex of a human to replicate the same style of thinking and perception. Neural Networks are organized in layers made up of interconnected nodes which contain an activation function that computes the output of the network.

We got the following best parameters through GridSearchCV for the dataset,

Neural Network Classifier(hidden_layer_sizes=**200**, max_iter=**1000**, tol=**0.01**)

Train Data:
AUC: 71%
Accuracy: 76%
Sensitivity: 57%
Precision: 65%
f1-Score: 61%

Test Data:
AUC: 71%
Accuracy: 77%
Sensitivity: 60%
Precision: 55%
f1-Score: 66%

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

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

Decision Tree Classifier:

Model Performance Evaluation on **Training data:**

Confusion Matrix:

```
array([[1215,  144],
       [ 308,  335]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.89	0.84	1359
1	0.70	0.52	0.60	643
accuracy			0.77	2002
macro avg	0.75	0.71	0.72	2002
weighted avg	0.77	0.77	0.76	2002

Model Performance Evaluation on **Testing data:**

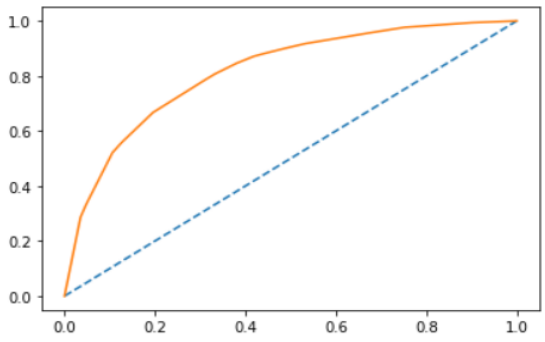
Confusion Matrix:

```
array([[531,  57],
       [133, 138]], dtype=int64)
```

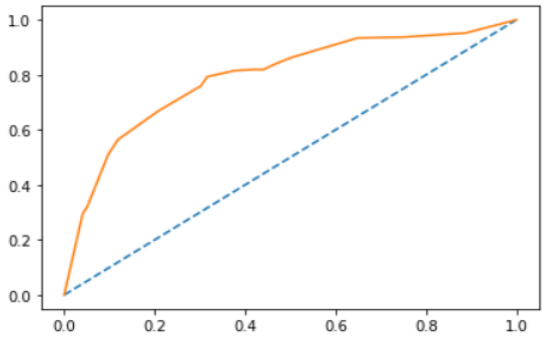
Classification Report:

	precision	recall	f1-score	support
0	0.80	0.90	0.85	588
1	0.71	0.51	0.59	271
accuracy			0.78	859
macro avg	0.75	0.71	0.72	859
weighted avg	0.77	0.78	0.77	859

ROC Curve



Train Data - AUC : 0.816



Test Data - AUC : 0.792

Random Forest Classifier:

Model Performance Evaluation on **Training data:**

Confusion Matrix:

```
array([[1192,  167],
       [ 263,  380]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.88	0.85	1359
1	0.69	0.59	0.64	643
accuracy			0.79	2002
macro avg	0.76	0.73	0.74	2002
weighted avg	0.78	0.79	0.78	2002

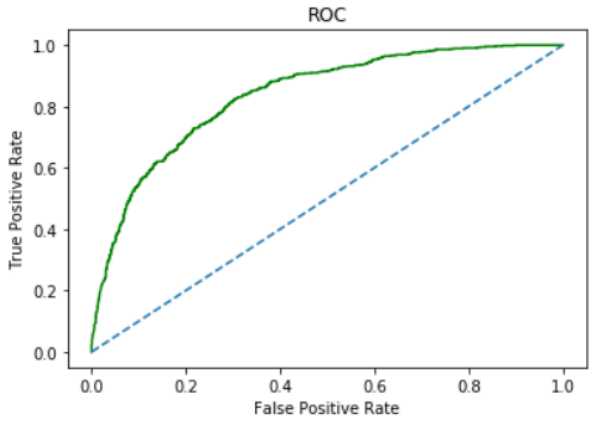
Model Performance Evaluation on **Testing data:**

Confusion Matrix:

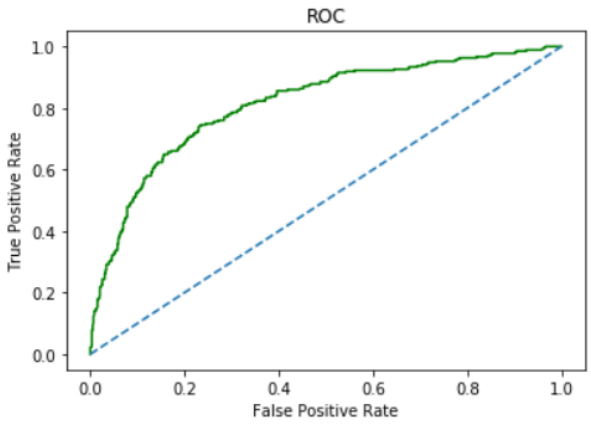
```
array([[516,  72],
       [114, 157]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.82	0.88	0.85	588
1	0.69	0.58	0.63	271
accuracy			0.78	859
macro avg	0.75	0.73	0.74	859
weighted avg	0.78	0.78	0.78	859



Train Data : AUC-0.734



Test Data : AUC -0.728

Neural Network Classifier:

Model Performance Evaluation on **Training data:**

Confusion Matrix:

```
array([[1159,  200],
       [ 275,  368]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.85	0.83	1359
1	0.65	0.57	0.61	643
accuracy			0.76	2002
macro avg	0.73	0.71	0.72	2002
weighted avg	0.76	0.76	0.76	2002

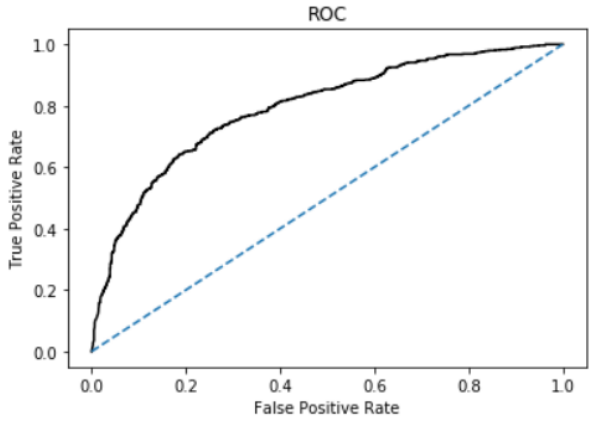
Model Performance Evaluation on **Testing data:**

Confusion Matrix:

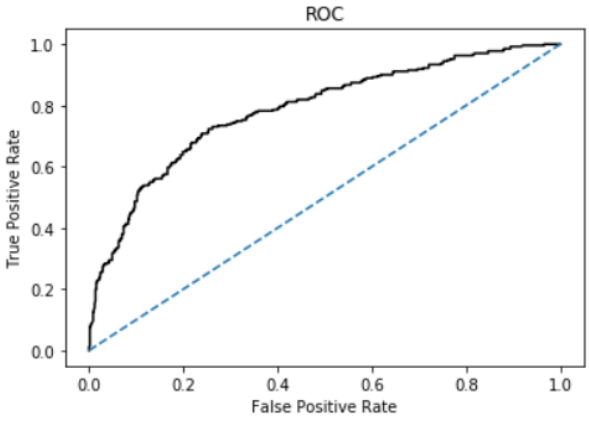
```
array([[512,  76],
       [122, 149]], dtype=int64)
```

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.87	0.84	588
1	0.66	0.55	0.60	271
accuracy			0.77	859
macro avg	0.73	0.71	0.72	859
weighted avg	0.76	0.77	0.76	859



Train Data : AUC-0.712



Test Data : AUC -0.710

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

Comparing the performance metrics from the three models, we can summarize as below,

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.77	0.78	0.79	0.78	0.76	0.77
AUC	0.82	0.79	0.73	0.73	0.71	0.71
Recall	0.52	0.51	0.59	0.58	0.57	0.60
Precision	0.70	0.71	0.69	0.69	0.65	0.55
F1 Score	0.60	0.59	0.64	0.63	0.61	0.66

Looking at the details got from **test data** from the three models ,

Accuracy : CART & Random Forest models have high value of 0.78

AUC : CART has highest value of 0.79 and Neural Network model has least value of 0.71

Recall : Neural Network model has highest value of 0.6 and CART model has least value 0.51

Precision : CART has highest value of 0.71 and Neural Network model has least value of 0.55

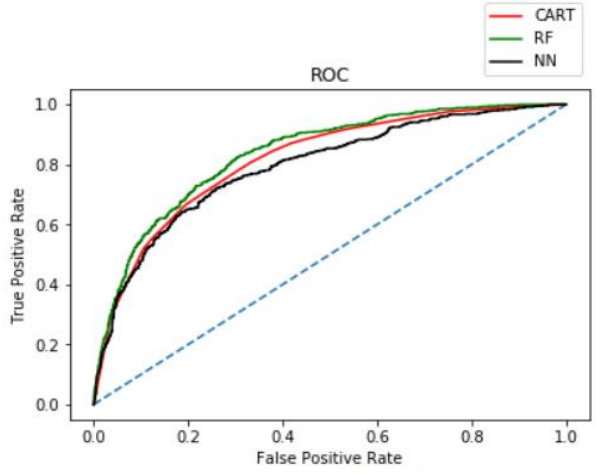
F1 Score : Neural Network has highest value 0.66 and CART model has least value of 0.59

Training and Test set results are almost similar in all the three models and overall measures are high in Random Forest.

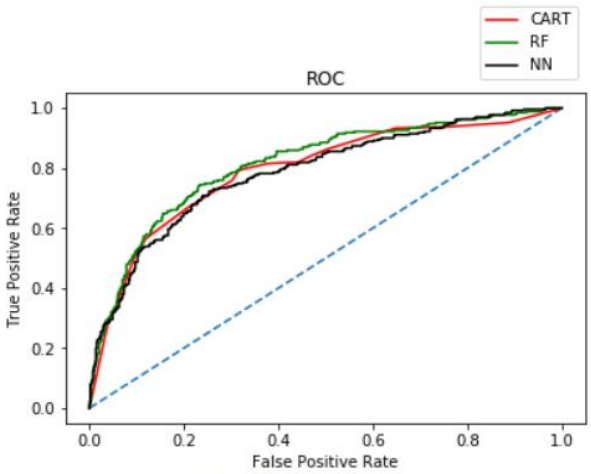
Therefore, **Random Forest has slightly better performance than the Cart and Neural network model**

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_Code** is found to be the most useful feature amongst all other features for predicting claim status.

Also Random forest has proven to be a great algorithm if the dataset is in tabular format. Random Forests requires less pre-processing and the training process is also much simpler. Moreover hyper-parameter tuning is easier with random forest when compared to neural networks. This gives random forest the edge above neural networks.



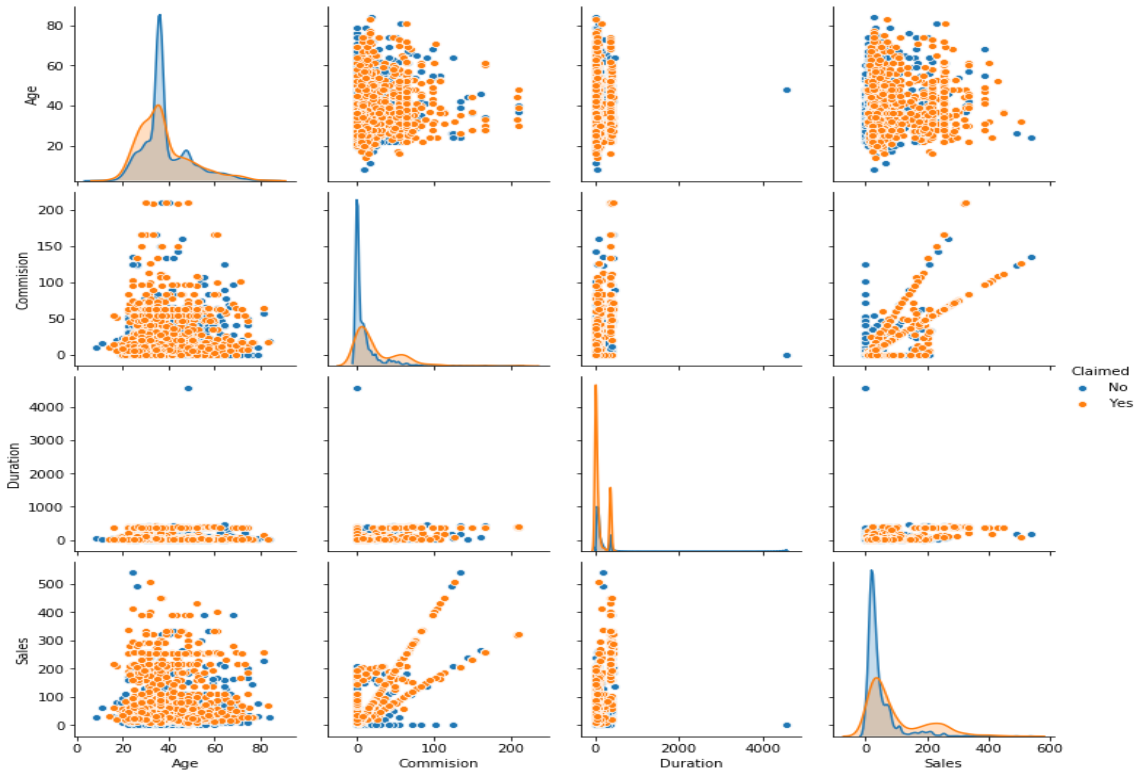
ROC Curve - Train Data



ROC Curve - Test Data

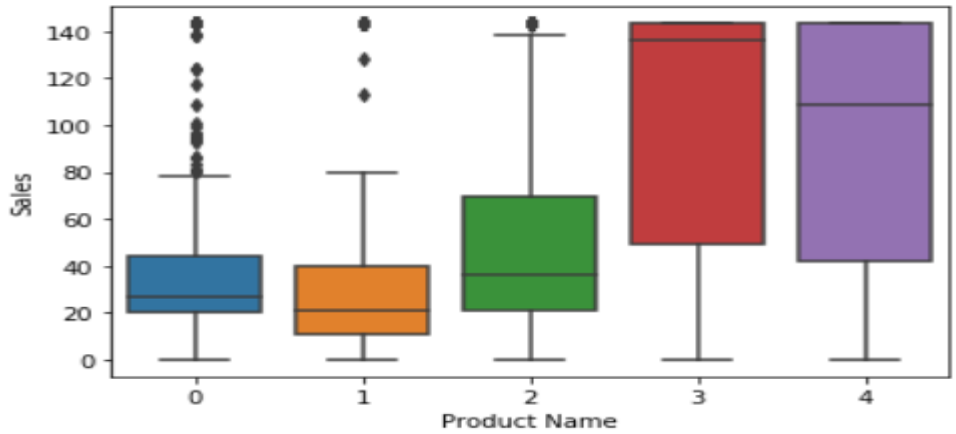
2.5) Inference: Basis on these predictions, what are the business insights and recommendations

Due to the importance of understanding and managing the risks in volatile business domains, it is required to find an effective aid in making decisions. The results from models show that Random Forest Trees algorithm is a promising opportunity in predicting claim status for the given data set.



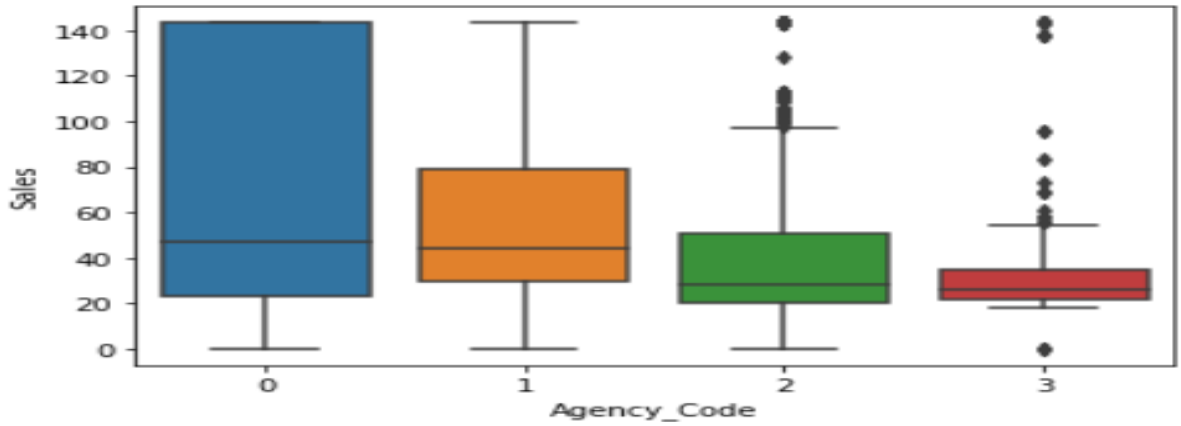
Claimed status 'No' & 'Yes' have values in similar range for all the variables 'Age', 'Commision','Duration' and 'Sales' .

Claimed status 'No' has highest magnitude in variables 'Age','Comission' & 'Sales' and Claimed status 'Yes' has highest magnitude in variable 'Duration'.



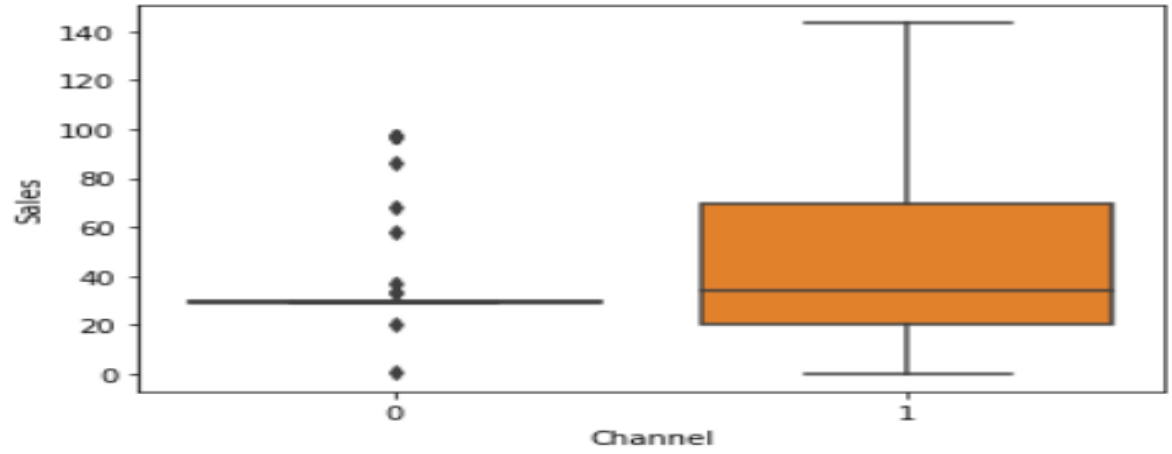
Product Name :Bronze Plan -0, Cancellation Plan-1, Customised Plan-2, Gold Plan-3, Silver Plan-4

From the above plot ,we can see Silver Plan has highest sales and Bronze Plan has least sales. Therefore certain measures have to be taken to increase sales through other plans also.



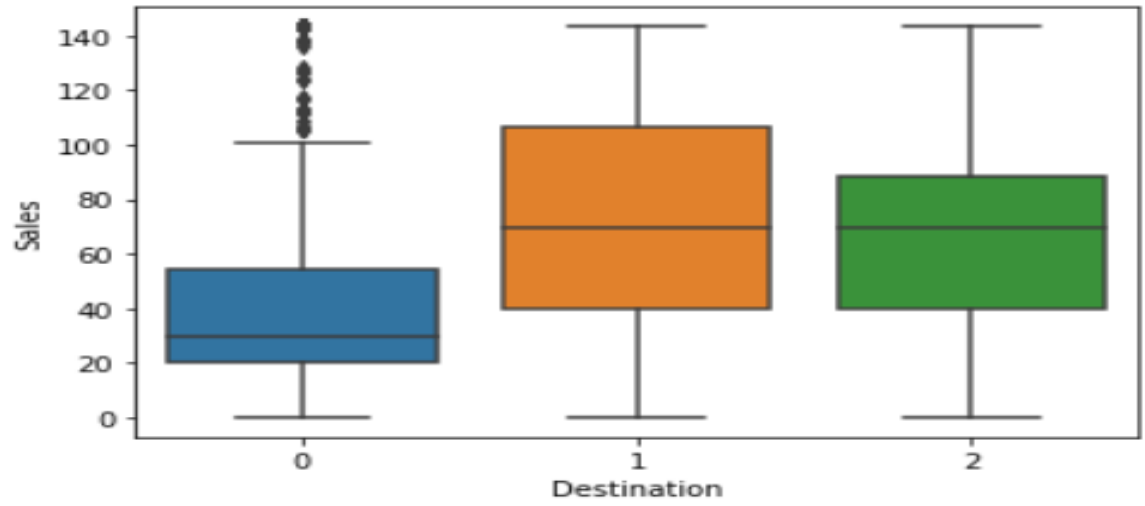
Agency_Code : C2B -0, EPX – 2,CWT-1,JZI-3

Sales is highest through tour firm C2B and least through tour firm JZI. Remaining agencies could also increase their sales by giving some offers to the customers.



Channel : Online -1, Offline– 0

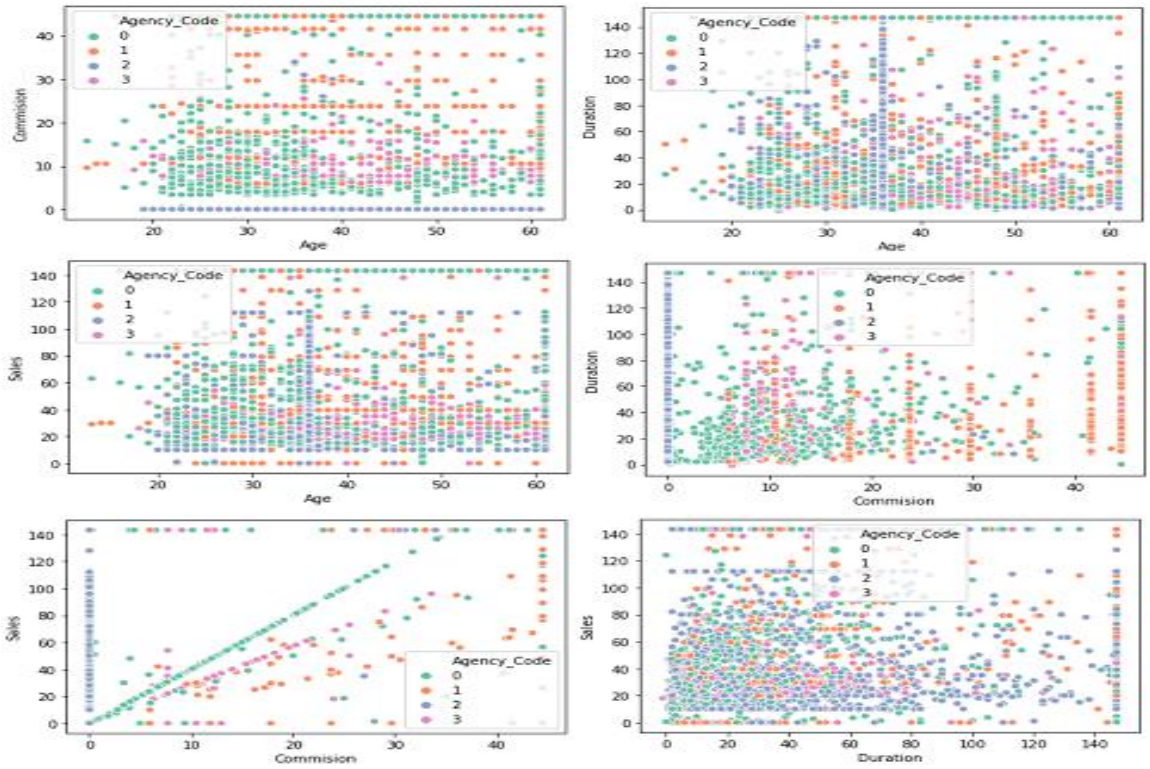
Tour insurance agencies have high proportion of distribution channels through Online , so offline channel distribution could be expanded to have higher reach to the customers.



Destination : ASIA -0, Americas– 1,EUROPE-2

Destination of the tour is higher for Americas and Europe and lower for Asia. So, Focusing on Asia destination would help for increase in sales.

Agency_Code is the most important variable for predicting claim status. So, looking at different variables with respect to variable 'Agency_Code', we have below plots



Agency_Code : C2B -0, EPX – 2,CWT-1,JZI-3

Sales is in high linear correlation with commission for Agency C2B & JZI and Sales is not affected by Commission for Agency CWT.

Duration is not affected by Commission for Agency CWT and Duration is increasing with higher Commission for Agency EPX.

From the overall results of the three models, the variable **Agency_code** is found to be the most useful feature amongst all other features for predicting the claim status.

Variable **'Sales'** is also important in deciding among different options available in other variables to get the best results for the Insurance firm.

So, The Overall analysis of given dataset definitely helped to get insights that would help the management for the business development.