DATA Mining Problem -2 Report

Submitted By :

Kanhaiya Awasthi

PGPDSBA SEP

Problem2:- 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.

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

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

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

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

Dataset for Problem 2: insurance_part2_data-1.csv

Attribute Information:

1. Target: Claim Status (Claimed)
2. Code of tour firm (Agency_Code)
3. Type of tour insurance firms (Type)
4. Distribution channel of tour insurance agencies (Channel)
5. Name of the tour insurance products (Product)
6. Duration of the tour (Duration)
7. Destination of the tour (Destination)
8. Amount of sales of tour insurance policies (Sales)
9. The commission received for tour insurance firm (Commission)
10. Age of insured (Age)

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

We are having Data set in CSV format named as insurance_part2_data-1.csv

- **Let's Have a Look on head of the Data:-**

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

- **Shape of Dataset:-** (3000,10) Dataset is having 3000 rows and 10 Columns
- **Information about the Data set:-**

```
<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
```

looking at Data info there are no Null values present. There are Only 4 Numerical Columns rest of them are Categorical and object type .

- Numerical Columns:-

- Age
- Commission
- Duration
- Sales

❖ Categorical Columns (Object or String):-
- Agency Code
- Type
- Claimed
- Channel
- Product Name
- Destination

As we Know that Most of Machine learning modal takes only Numerical and Ca tegorical Data Only (Categorical Data should be converted into ordinal form). S o we convert all the columns into Ordinal or numerical form various encoding t echniques.

➢ **Descriptive Statistics of the Data:-**

Using describe function we found the table as below

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
count	3000	3000	3000	3000	3000	3000	3000	3000	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	38.09	NaN	NaN	NaN	14.53	NaN	70.00	60.25	NaN	NaN
std	10.46	NaN	NaN	NaN	25.48	NaN	134.05	70.73	NaN	NaN
min	8.00	NaN	NaN	NaN	0.00	NaN	-1.00	0.00	NaN	NaN
25%	32.00	NaN	NaN	NaN	0.00	NaN	11.00	20.00	NaN	NaN
50%	36.00	NaN	NaN	NaN	4.63	NaN	26.50	33.00	NaN	NaN
75%	42.00	NaN	NaN	NaN	17.24	NaN	63.00	69.00	NaN	NaN
max	84.00	NaN	NaN	NaN	210.21	NaN	4580.00	539.00	NaN	NaN

Age Ranges from 8 to 84 years & commission varies from 0 to 210.10 also note that Claim, Type and channel having 2 unique values Agency code, Product Nam e & destination having 4,5,3 unique values respectively.

- **Checking of Duplicate Records: -** By executing Dupes function in the data set we found that there are 139 duplicate rows present in data So one st rategy can be removing the duplicate rows from the data **but there is no unique Identity of customers present** so this data can be of different cus tomers of an Insurance company, due to this uncertainty We are not rem oving the Duplicate rows .

- **Checking Distribution of Numerical columns visually: -**

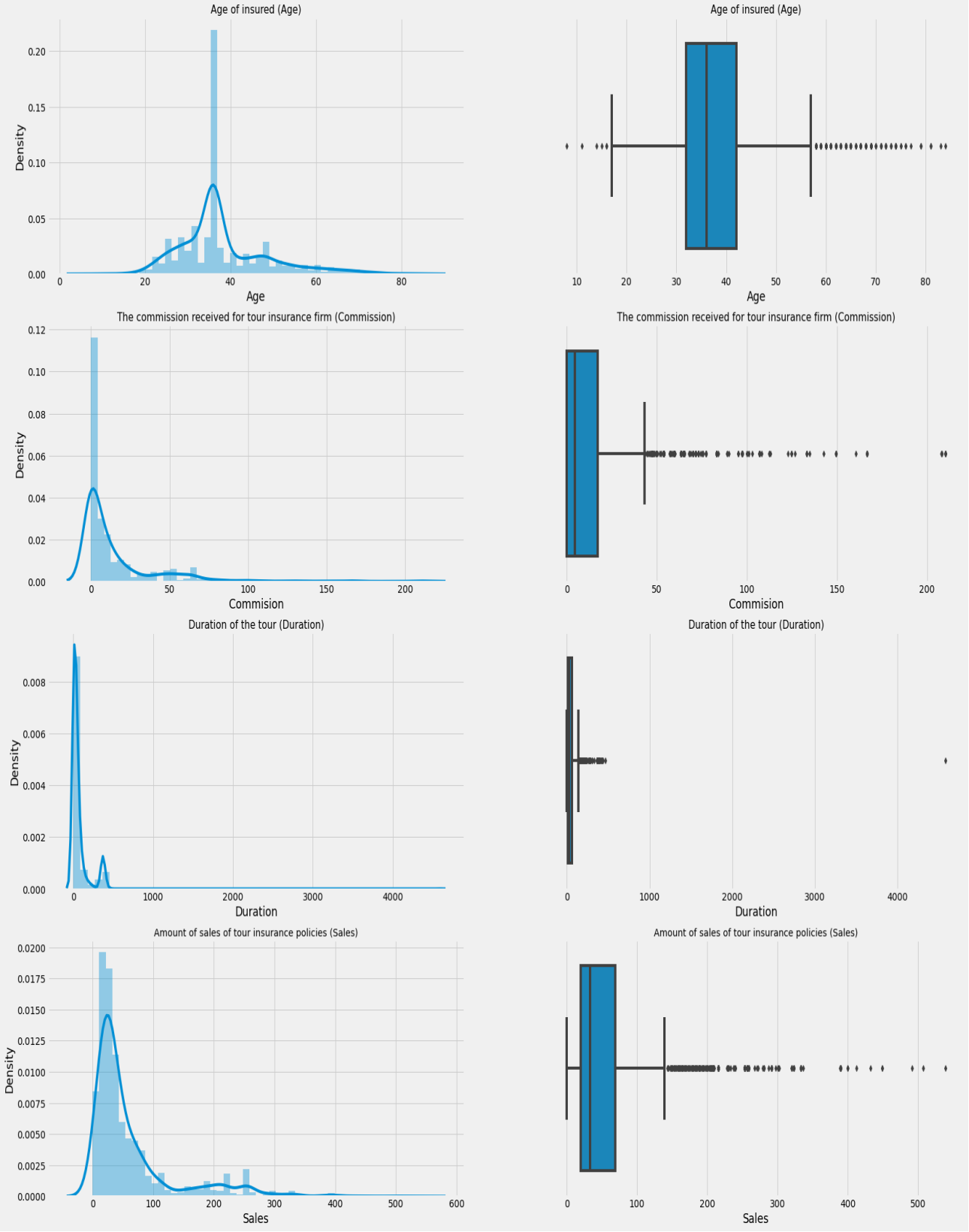


Fig:-1.1

In above figure We have made distribution plot for all the four Numerical as bel ow

- Numerical Columns:-
  - Age
  - Commission
  - Duration
  - Sales

Referring to boxplot and distplot We can clearly see that Numerical data points are left skewed and having many outliers . So we have to treat Outliers first (Bu t in the Quiz FAQs it is said that Outlier treatment is not necessary) .

Finding patterns in Categorical (Object or String) Data V isually:-

- Categorical Columns (Object or String):-
  - Agency Code
  - Type
  - Claimed
  - Channel
  - Product Name
  - Destination

Categorical data can be best represented by frequency plot boxplot as well as V iolin plots so we are going to represent data using all the 3 plots

We will be plotting Boxplot and Violin plot between Agency & corresponding sa les of that agency and Claim status as Hue that will be telling us how much amo unt of Sales made by that Agency and Out of them how many Customers have c laimed the insurance.

Below are the Unique Values of categorical Columns and their frequency

```
Agency_Code
EPX    1365
C2B     924
CWT     472
JZI     239
Name: Agency_Code, dtype: int64


Type
Travel Agency    1837
Airlines         1163
Name: Type, dtype: int64



Claimed
 No     2076
```

```
Yes      924
Name: Claimed, dtype: int64
```

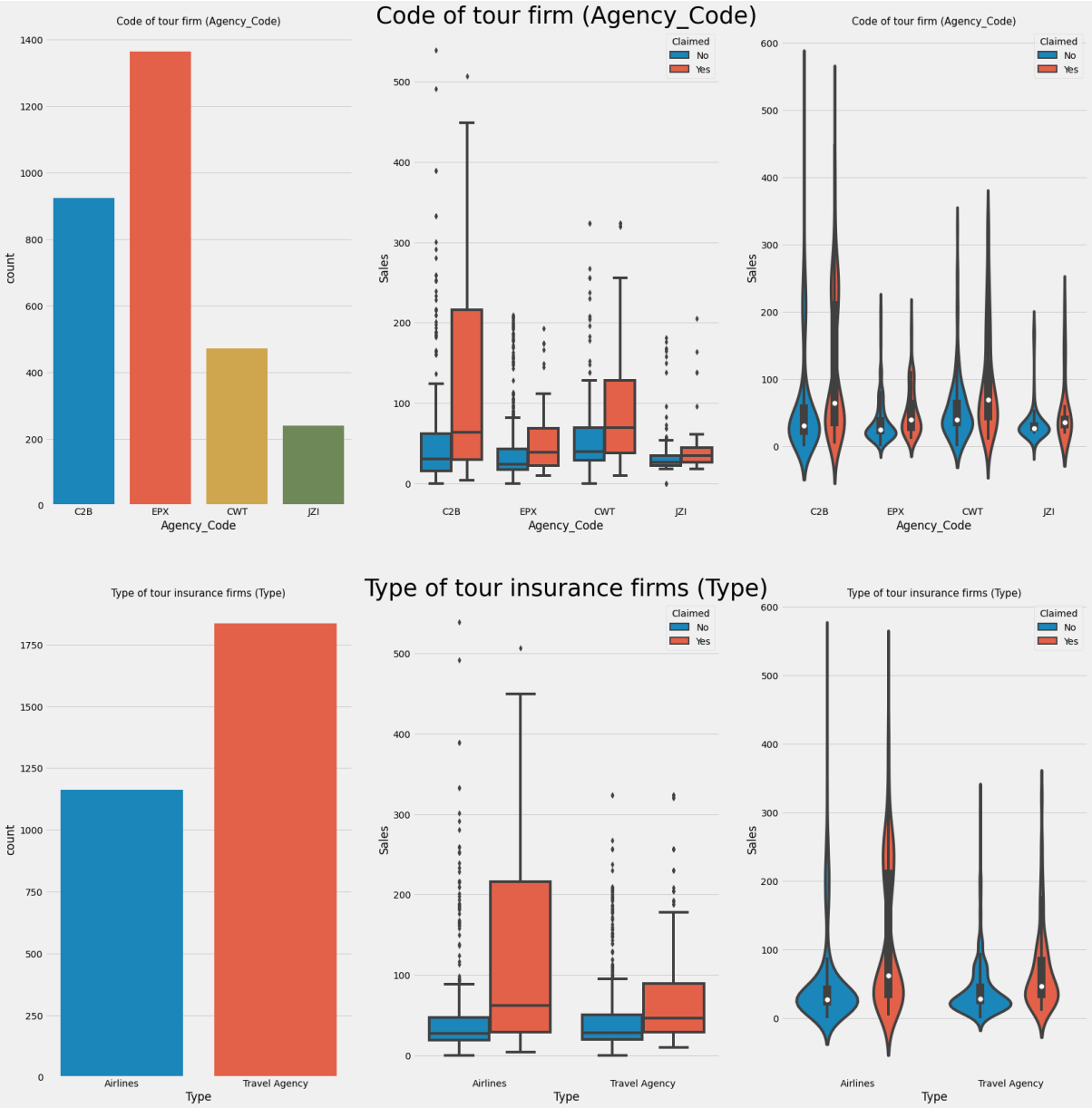


Fig- 1.2

We can clearly see from above plot that Agency code EPX having highest frequency i.e. 1365 that means they are having more number of customer transactions & JZI having lowest frequency i.e. 239 .You can also note from Boxplot that Agency which is having higher mean sales their insurance claims are high .

Below unique value Counts suggests us that more than 90% of the peoples cho
osing online channels for taking insurance

```
Channel
Online     2954
Offline      46
Name: Channel, dtype: int64
```

```
Product Name
Customised Plan      1136
Cancellation Plan     678
Bronze Plan           650
Silver Plan           427
Gold Plan             109
Name: Product Name, dtype: int64
```

```
Destination
ASIA         2465
Americas      320
EUROPE        215
Name: Destination, dtype: int64
```

Below Figure 1.3 tells us that 1136 customers have chosen Customized plan for
their tour and Very few have chosen Gold Plan for their Travel but median valu
e of Sales in Gold plan is High. Also Violin plot told us that most of the sales of T
our products are lying near their Median Values.

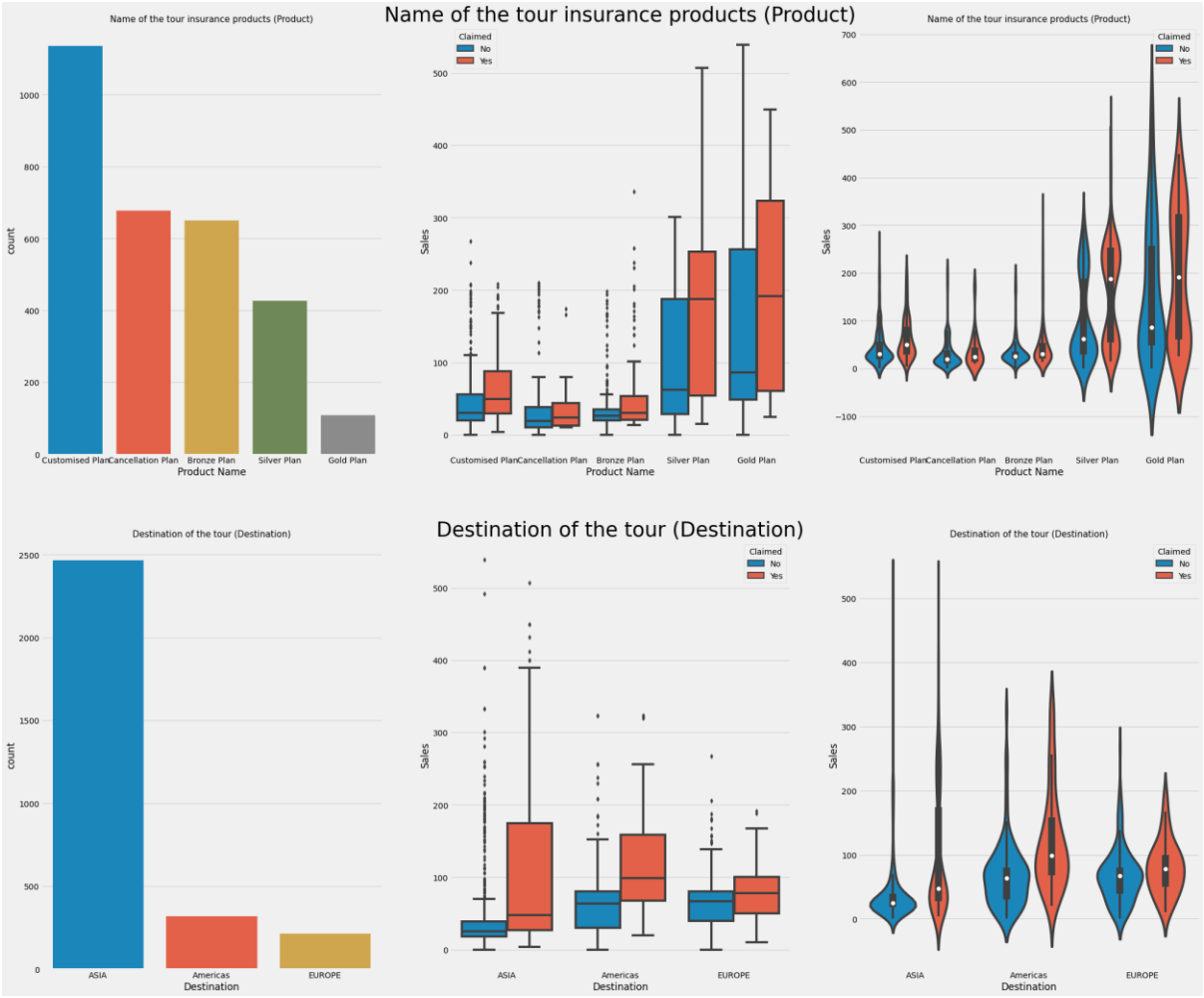
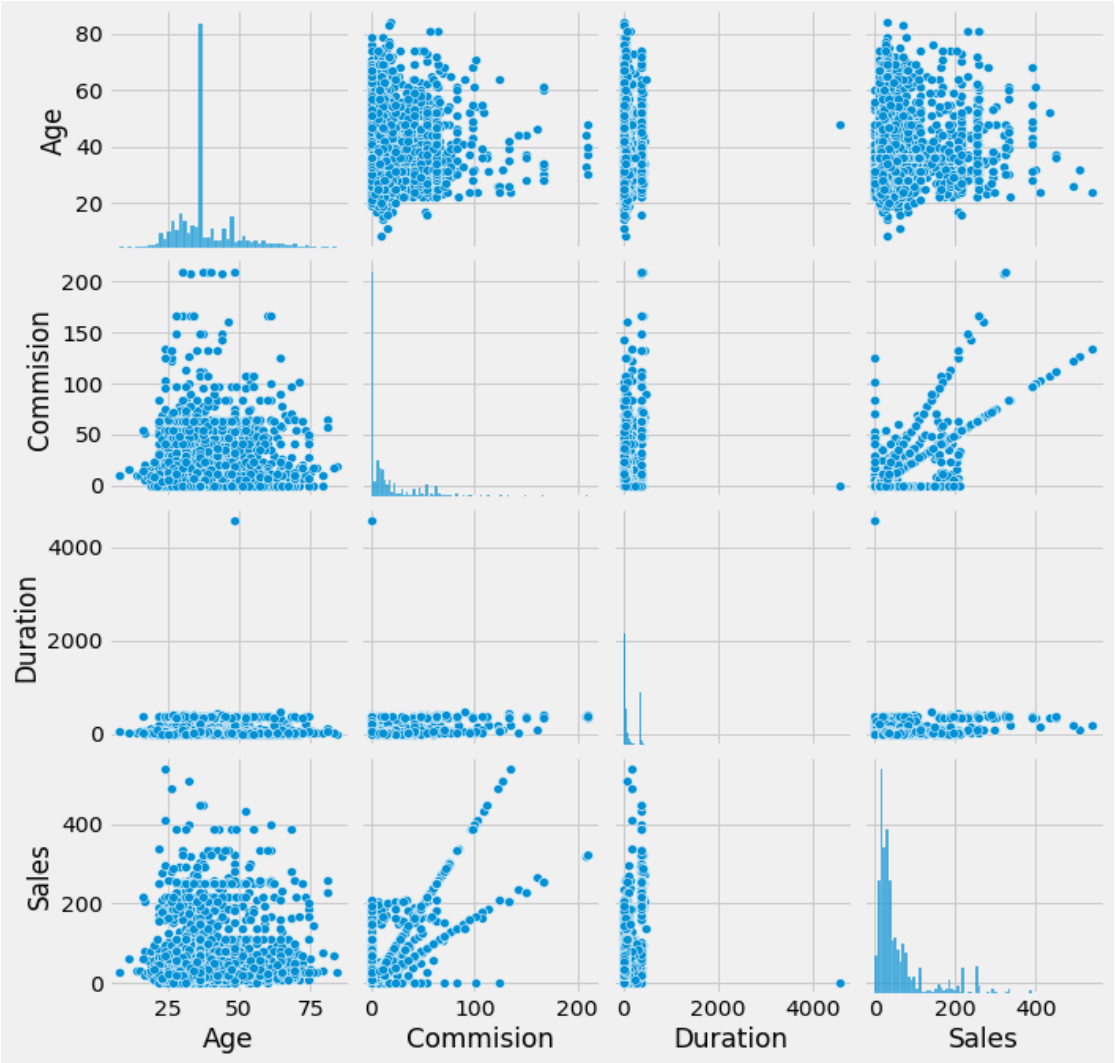

Name of the tour insurance products (Product)
count
Product Name
Customised Plan
Cancellation Plan
Bronze Plan
Silver Plan
Gold Plan
Name of the tour insurance products (Product)
Claimed
No
Yes
Sales
Product Name
Name of the tour insurance products (Product)
Claimed
No
Yes
Sales
Product Name
Destination of the tour (Destination)
count
Destination
ASIA
Americas
EUROPE
Destination of the tour (Destination)
Claimed
No
Yes
Sales
Destination
Destination of the tour (Destination)
Claimed
No
Yes
Sales
Destination

Fig: - 1.3

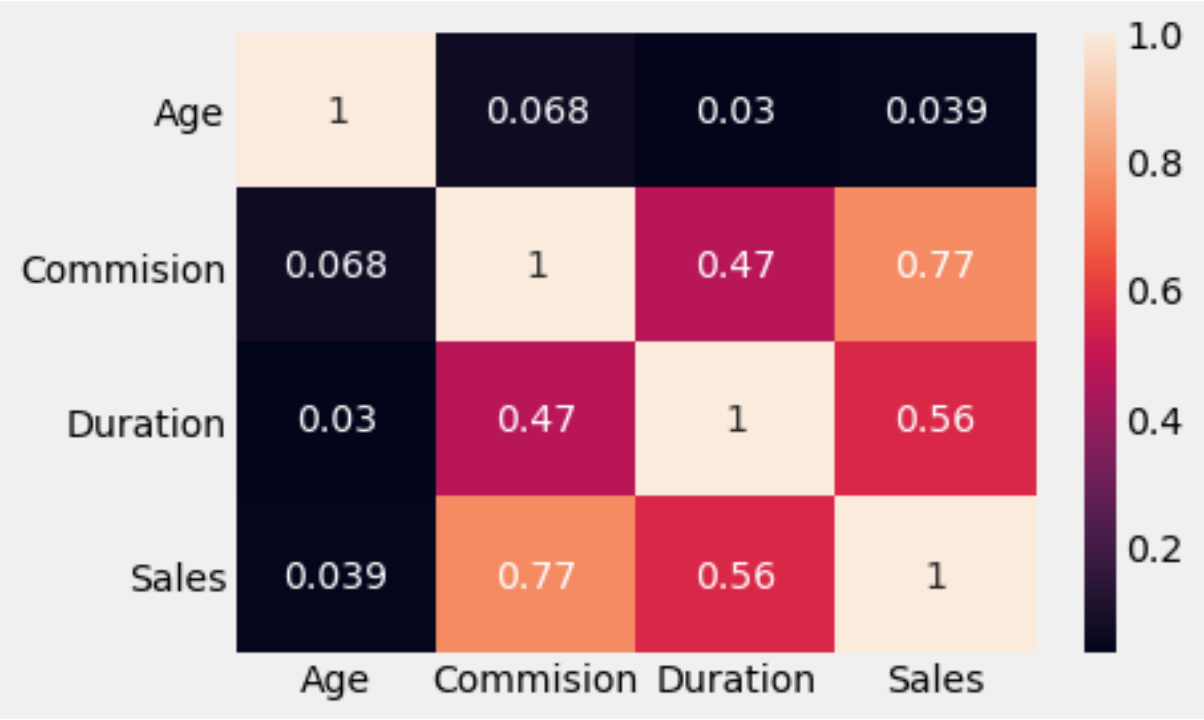
- **Now We will check patterns in the Data using Pair plot.**



Looking at the pair plot we found that there are high Amount of sales between Age bracket of 25 to 75 . Commission is also high as the sales is going High.

- **Checking Correlations Using Pearson Correlation Matrix :-**

	Age	Commision	Duration	Sales
Age	1.00	0.07	0.03	0.04
Commision	0.07	1.00	0.47	0.77
Duration	0.03	0.47	1.00	0.56
Sales	0.04	0.77	0.56	1.00



Looking at above correlation Matrix we can notice that Commission and Sales a re highly Correlated & Duration and Sales are also significantly correlated & all other Data have very less Correlation.

- **Since Decision Tree ,CART & ANN in Python can take only numerical / c ategorical columns. It cannot take string / object types.**
  So we have change Categorical Columns into numerical column by chang ing their data types and Found below results

```
feature:
Agency_Code
[C2B, EPX, CWT, JZI]
Categories (4, object): [C2B, CWT, EPX, JZI]
```

```
[0 2 1 3]


feature: Type
[Airlines, Travel Agency]
Categories (2, object): [Airlines, Travel Agency]
[0 1]


feature: Claimed
[No, Yes]
Categories (2, object): [No, Yes]
[0 1]


feature: Channel
[Online, Offline]
Categories (2, object): [Offline, Online]
[1 0]


feature: Product Name
[Customised Plan, Cancellation Plan, Bronze Plan, Silver Plan, Gold Pla
n]
Categories (5, object): [Bronze Plan, Cancellation Plan, Customised Pla
n, Gold Plan, Silver Plan]
[2 1 0 4 3]


feature: Destination
[ASIA, Americas, EUROPE]
Categories (3, object): [ASIA, Americas, EUROPE]
[0 1 2]
```

- **Let's have a look on Head of encoded Data**

Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
48	0	0	0	0.7	1	7	2.51	2	0
36	2	1	0	0	1	34	20	2	0
39	1	1	0	5.94	1	3	9.9	2	1
36	2	1	0	0	1	4	26	1	0
33	3	0	0	6.3	1	53	18	0	0

By doing this encoding we found that around 69.2 % of peoples haven't claimed insurance & only 30.8 % have claimed Insurance.

```
0 -   0.692

1  -  0.308
Name: Claimed, dtype: float64
```

%0s - 2076 Who haven't claimed Insurance

%1s -924 Who have claimed Insurance

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

First of all we split Data into dependent Variable Y (Claimed)and independent variable X .

Now We have to split the data into train and test set by using following parameters (X, y, test_size=.30, random_state=1) train data is 70% and test Data is 30% .
Training Data is having 2,100 rows and 9 columns , whereas test Data consist of 9,00 rows and 9 columns out of total 3,000 observations.

X_train (2100, 9)
X_test (900, 9)
train_labels (2100,)
test_labels (900,)
Total Obs 3000

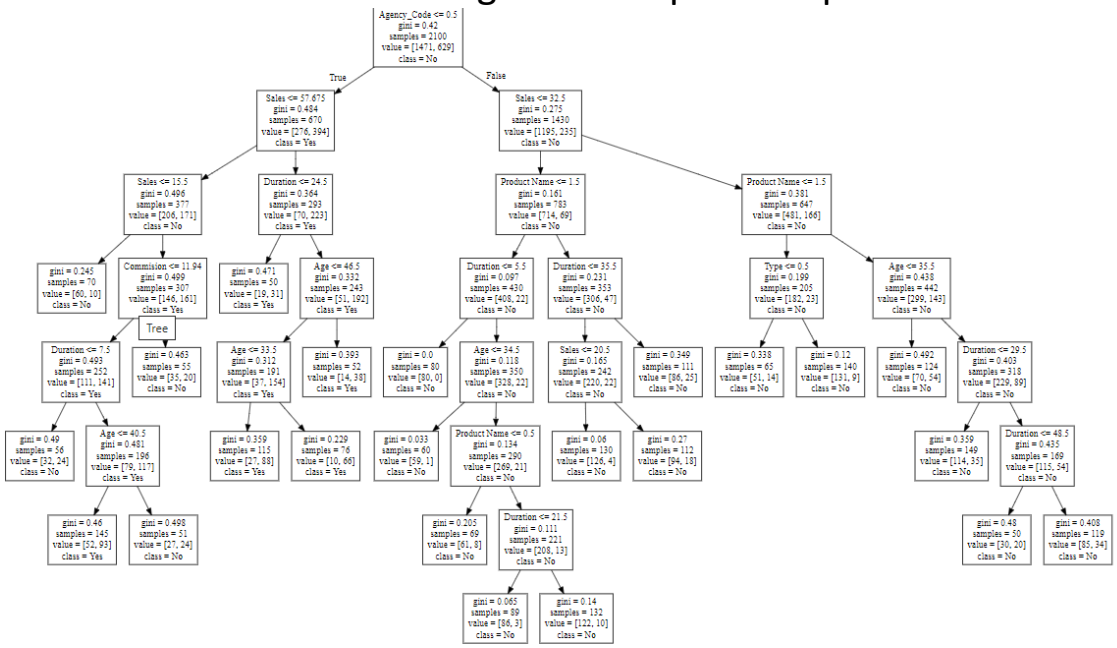
Build classification model CART, Random Forest, Artificial Neural Network

After Splitting the Data into Training and test set Our next step is to Build a classification Modal for **CART, Random Forest, Artificial Neural Network** by using various submodules of sklearn package library.(Refer Attached both jupyter Notebooks for Details)

- **CART :-** In Classification and Regression Technique (CART) we have made a modal by fitting train and test Data directly to DecisionT

reeClassifier and checked the Result Using http://webgraphviz.com/ a s we have seen that data gets over fitted . To encounter this prob lem, we will pass various parameters to Grid search CV to find out the best result.

- Below we have given a snapshot of pruned decision Tree



- **Random Forest :-** Random forest Modal is fitted into Train and test data & we have also used Grid search Cross Validation (Grids earchcv) function from python to find out best parametrs (Refer a ttached Jupyter Notebook for details)

- **Artificial Neural Networks** :-
- while creating ANN modal first we have scaled the training and te st Data using standard Scaling function.
- We have fitted Training and test data into Multi- layer perceptron Classifier (MLPClassifier)
- By using Grid search CV we find out the best parameters

❖ By finding out best parameters using Cross validation we will create class ification report confusion matrix and Accuracy score.

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**

We have Got Performance matrices (Confusion Matrix, Classification Report ROC curve, AUC Score) as below for All the three Modals :-

- **Decision Tree(Confusion Matrix, Classification Report ROC curve, AUC Score):-**

DT- Classification report for Train Dataset

	precision	recall	f1-score	support
0	0.81	0.92	0.86	1471
1	0.72	0.50	0.59	629
accuracy			0.79	2100
macro avg	0.77	0.71	0.73	2100
weighted avg	0.78	0.79	0.78	2100

DT--Classification report for Test Dataset

	precision	recall	f1-score	support
0	0.76	0.93	0.83	605
1	0.73	0.38	0.50	295
accuracy			0.75	900
macro avg	0.74	0.66	0.67	900
weighted avg	0.75	0.75	0.72	900

In training set we have Got the accuracy of 79 % and on test set accuracy is 75 % that means there is no significant drop of accuracy.

We are also concerned about Sensitivity first know what is sensitivity ?
Sensitivity is the ratio between how much were correctly identified as positive to how much were actually positive. Remember false ne gatives are data points which should be identified as true.

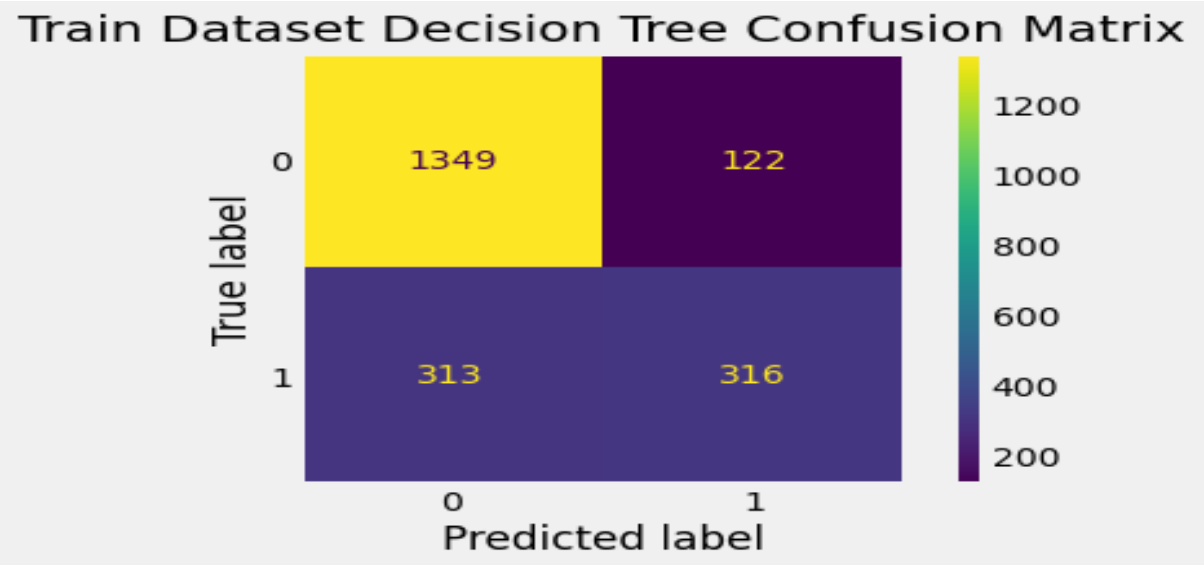
$$Sensitivity\ or\ Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Since Insurance Company is struggling to correctly predict Claim Status as yes s o we will take Claim status 1 (Yes) in our analysis.

We have made a crosstab below for better understanding Cart Train Modal

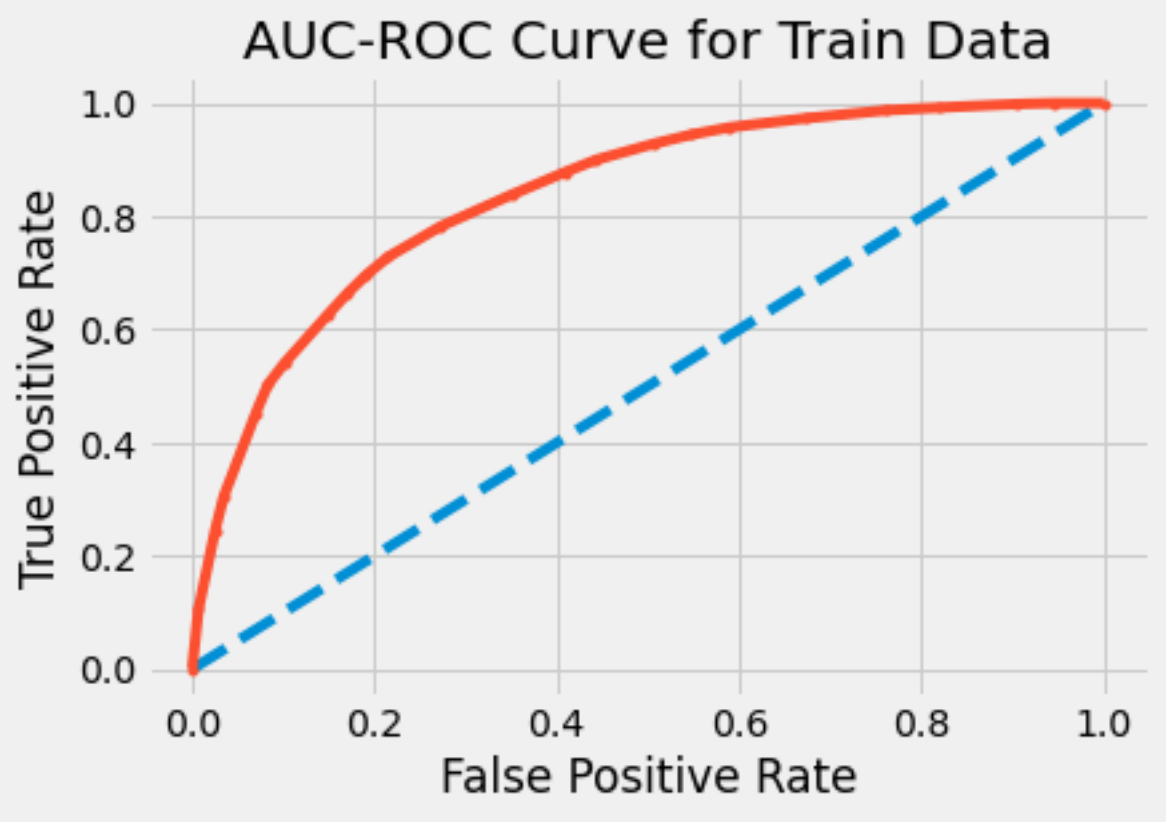
	CART Train Set	CART Test set
Accuracy	0.79	0.75
AUC	0.83	0.8
Recall	0.5	0.38
Precision	0.72	0.73
F1 Score	0.59	0.5

- Decision Tree Train Dataset Confusion Matrix
  In below confusion Matrix of Train dataset True positive are 313 while as 316 data points are False Negative .



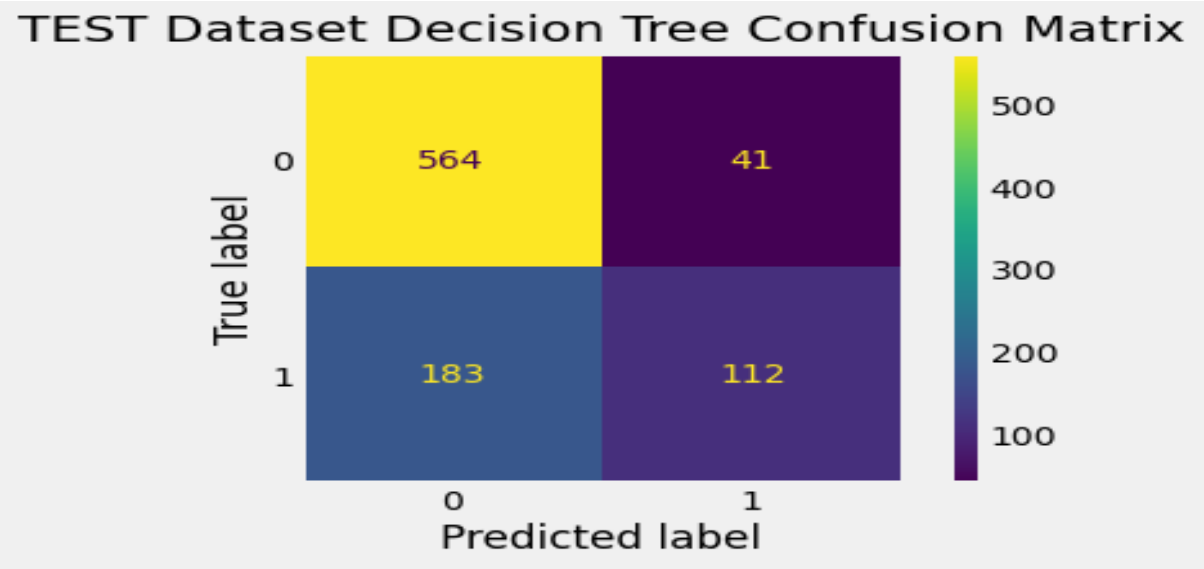
Decision Tree Train DATA AUC :- 83 %

- Roc Curve for Train Data is given below

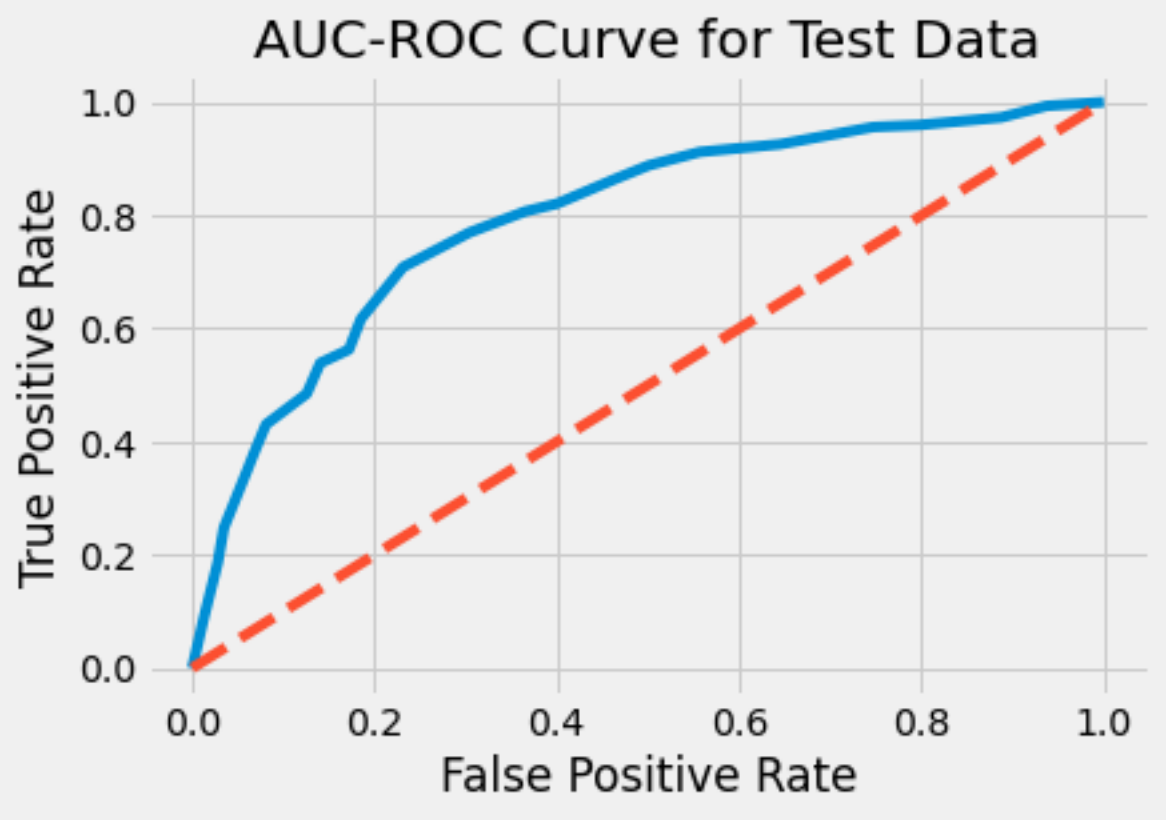


Decision Tree Performance evaluation Of test Data:-

- **Test Data Confusion Matrix:-**



- **Decision Tree AUC OF Test Data:- 79.4%**



Random Forest performance Evaluation: -

Just like Decision Tree we are making performance evaluation of Trai n and Test Data using Random Forest Modal .

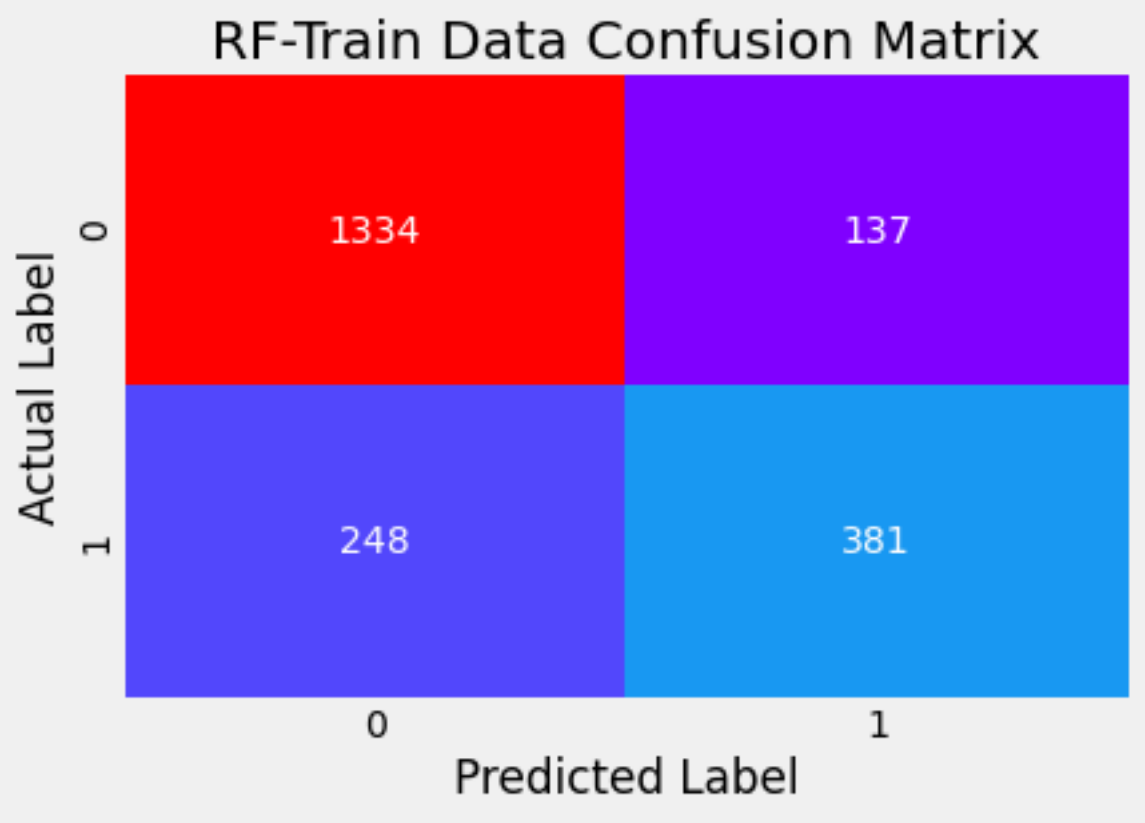
Random Forest Classification report for Train Dataset

	precision	recall	f1-score	support
0	0.84	0.91	0.87	1471
1	0.74	0.61	0.66	629
accuracy			0.82	2100
macro avg	0.79	0.76	0.77	2100
weighted avg	0.81	0.82	0.81	2100

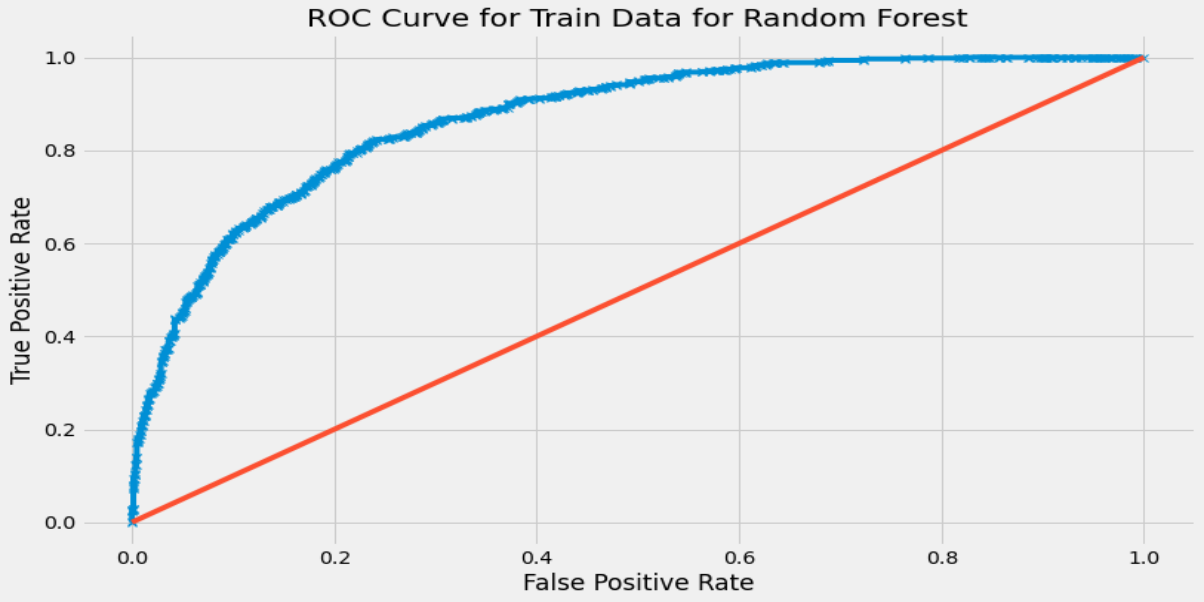
Random Forest Classification report for Test Dataset

	precision	recall	f1-score	support
0	0.78	0.91	0.84	605
1	0.73	0.48	0.58	295
accuracy			0.77	900
macro avg	0.76	0.70	0.71	900
weighted avg	0.77	0.77	0.76	900

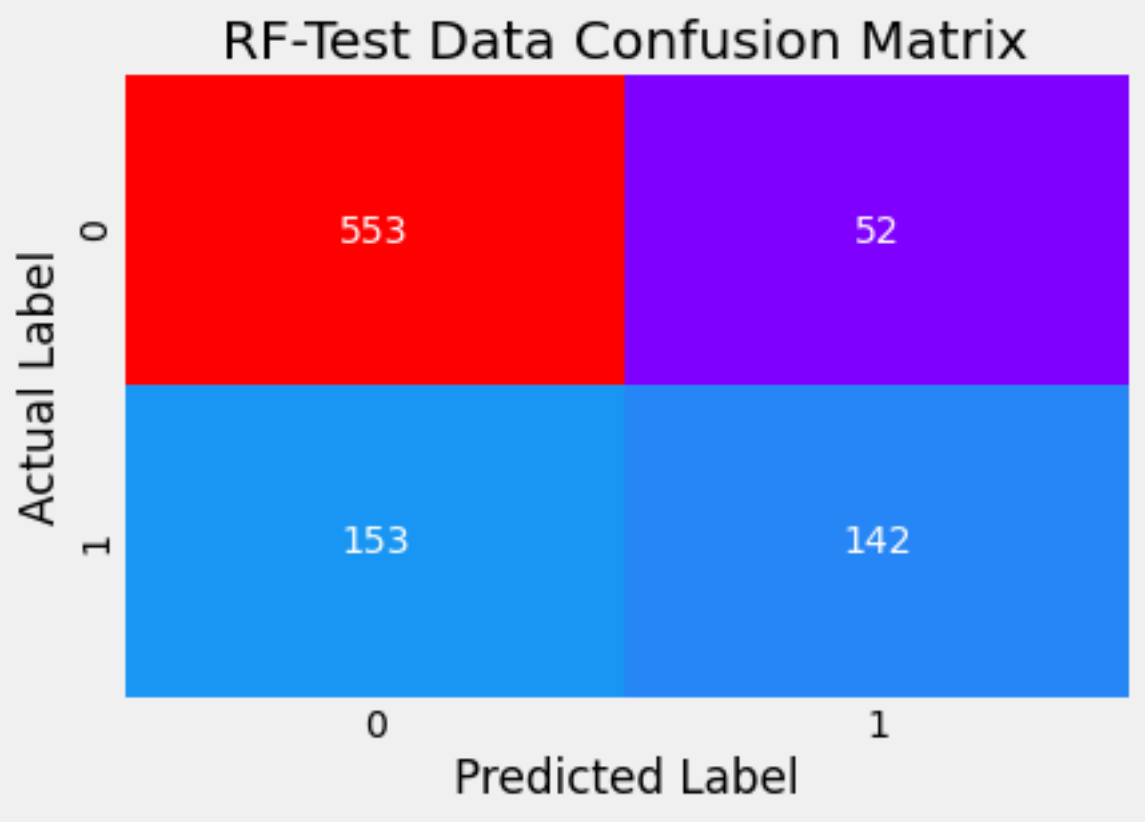
Random Forest Train Data Evaluation: -



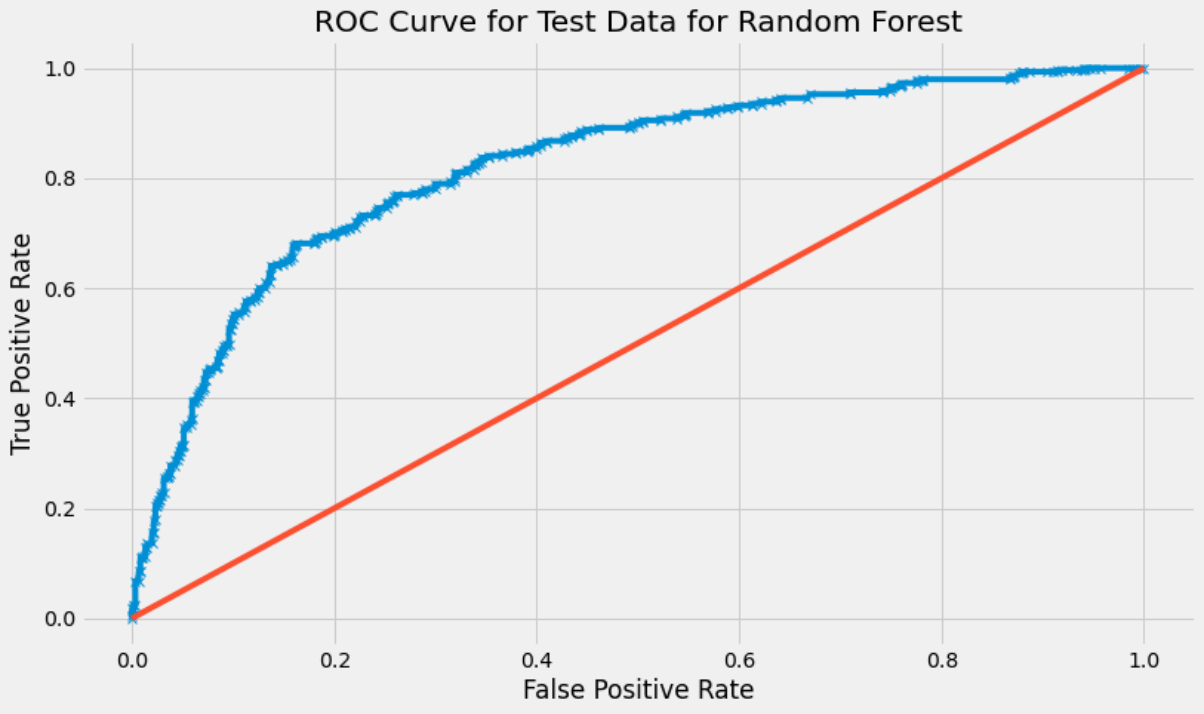
Random Forest Modal - AUC SCORE FOR TRAIN DATA :- 87%



- **Random Forest - Test Data Evaluation:-**



RANDOM FOREST AUC score for Test Data :- 83%



By analysing the above random forest Model, we find out the following results

	Random Forest Train Set	Random Forest Test Set
Accuracy	0.82	0.77
AUC	0.87	0.83
Recall	0.61	0.48
Precision	0.74	0.73
F1 Score	0.66	0.58

By looking at above Crosstab Train and Test dataset results are almost similar s o This modal is A Good Model.

Artificial Neural Networks Model:-

By passing Various values to Gridsearch to extract best parameters w e found out below resuts.

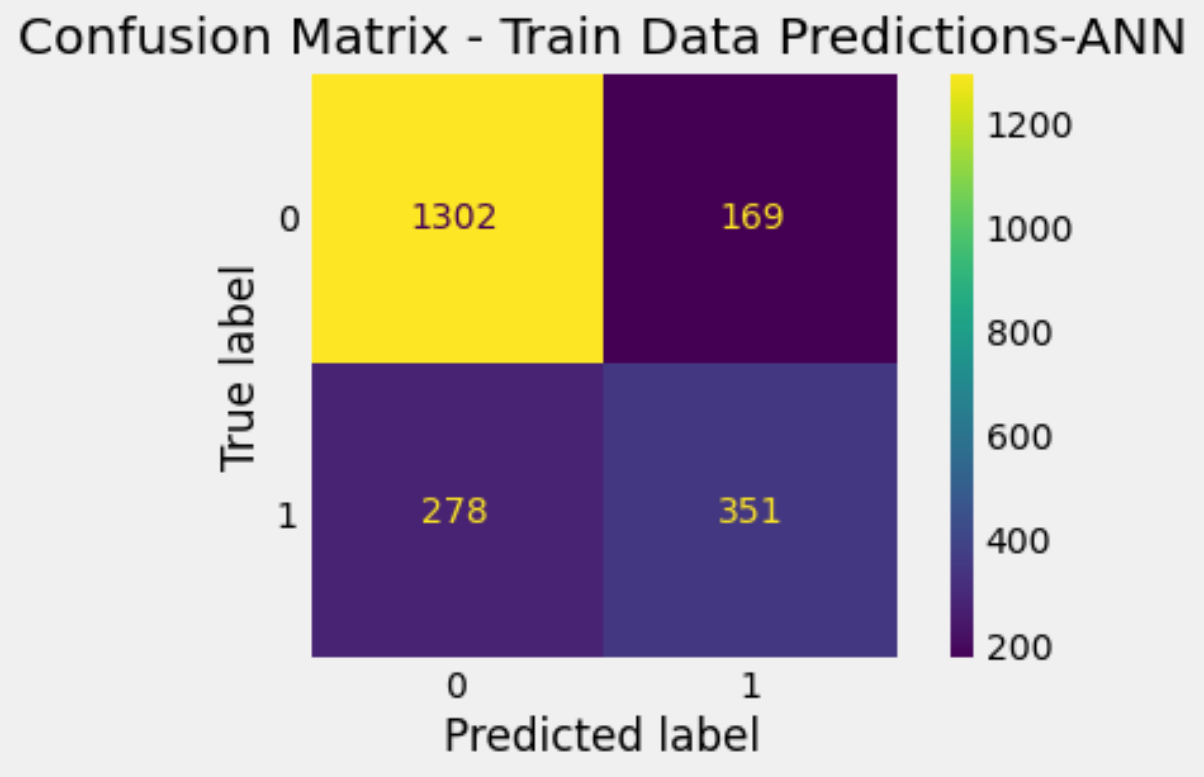
A--Classification report for Train Dataset for Hypertuned ANN model

	precision	recall	f1-score	support
0	0.82	0.89	0.85	1471
1	0.68	0.56	0.61	629
accuracy			0.79	2100
macro avg	0.75	0.72	0.73	2100
weighted avg	0.78	0.79	0.78	2100

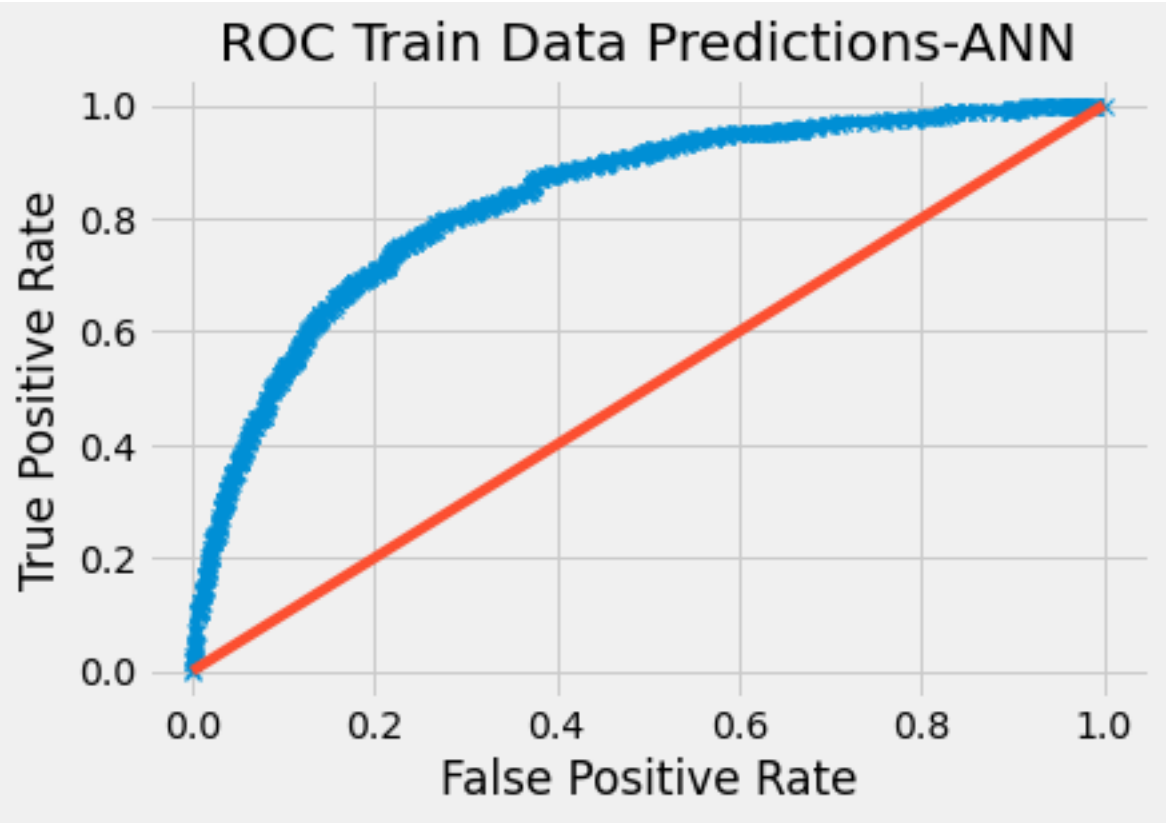
B--Classification report for Test Dataset for Hypertuned ANN model

	precision	recall	f1-score	support
0	0.78	0.91	0.84	605
1	0.72	0.46	0.56	295
accuracy			0.77	900
macro avg	0.75	0.69	0.70	900
weighted avg	0.76	0.77	0.75	900

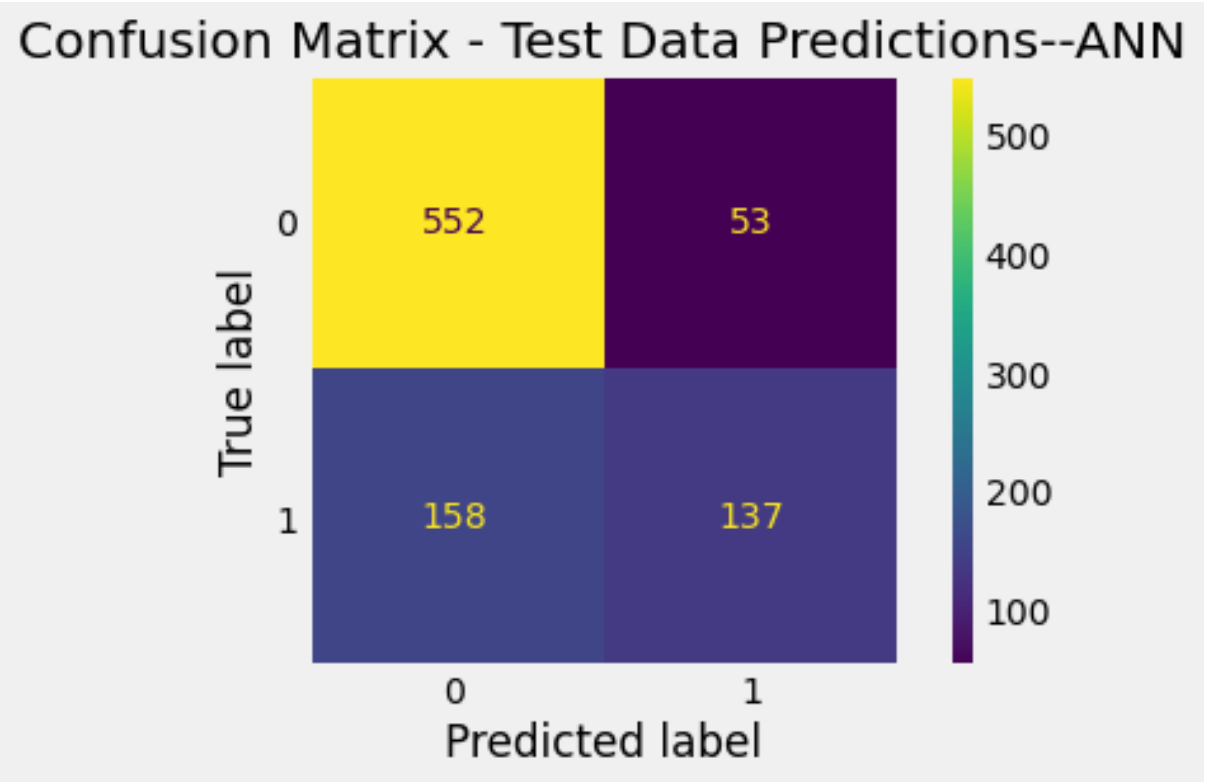
- **ANN Model Train Dataset Prediction:-**



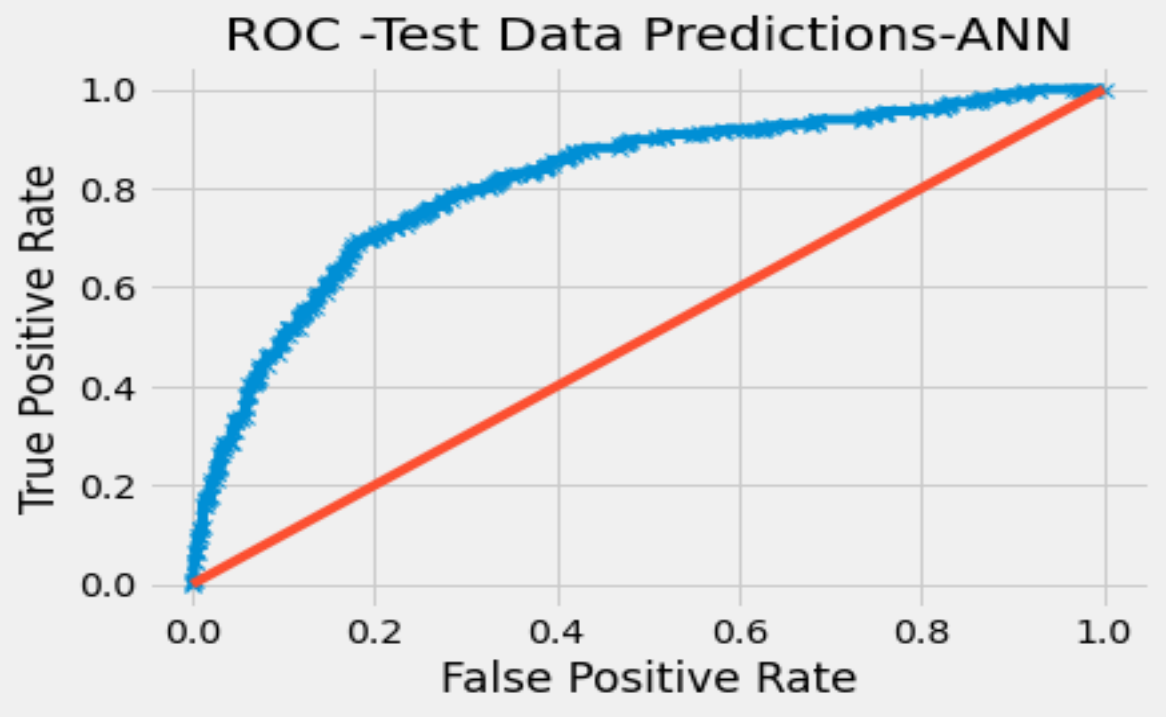
ARTIFICIAL NEURAL NETWORK AUC Score of Train data:- 84%



- **ANN Model Test Dataset Prediction:-**



ANN AUC Score of Test Data:- 82%



By analysing all of the above Model we found out below results

	Neural Network Train set	Neural Network Test set
Accuracy	0.79	0.77
AUC	0.84	0.82
Recall	0.56	0.46
Precision	0.68	0.72
F1 Score	0.61	0.56

Looking at above table we found that Train and test Data set results are almost similar so This Model is a good Model.

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

Evaluating all the three Models (CART, Random Forest & Neural Networks) we h ave found below Results

	CART Train Set	CART Test set	Random Forest Train Set	Random Forest Test Set	Neural Network Train set	Neural Network Test set
Accuracy	0.79	0.75	0.82	0.77	0.79	0.77
AUC	0.83	0.8	0.87	0.83	0.84	0.82
Recall	0.5	0.38	0.61	0.48	0.56	0.46
Precision	0.72	0.73	0.74	0.73	0.68	0.72
F1 Score	0.59	0.5	0.66	0.58	0.61	0.56

As we can see that in all the models Train and Test results are almost similar bu t we have to compare all the Models and find out the best model among all.

Having a Glance on the table we can noticed Green Colour for Random Forest M odel which shows that Train and test set results are high among all the models s o can Conclude that Random forest model is best/optimized Model.

Neural network Model have given better results but their results are lesser opti mized in comparison to Random Forest Model. Thus this model is medium perf ormed model.

As you can also see red marked in the report of CART test in which both Train a nd Test set results are Low as compared to all so this is Lowest Optimized Mode l.

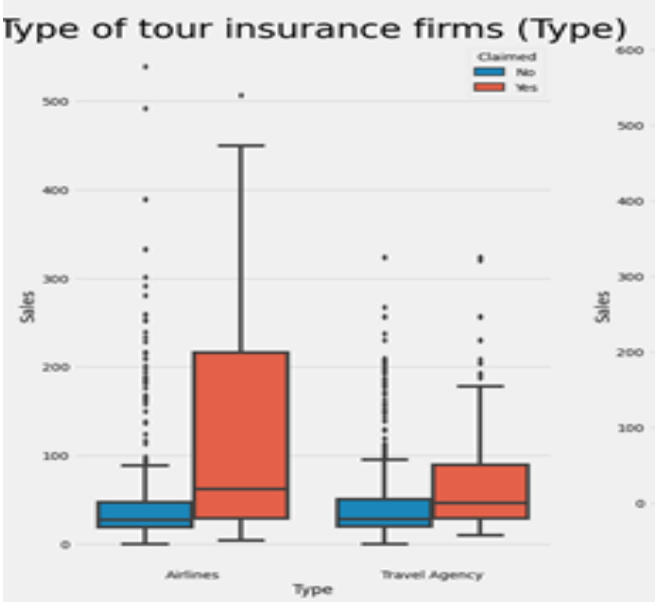
2.5 Based on your analysis and working on the business problem, de tail out appropriate insights and recommendations to help the man agement solve the business objective.

By using Various Algorithms like CART, Random Forest and Artificial Neural Net works we analysed the data and found that More data needs to be collected lik e there is no unique identifier present so we are unable to clearly distinguished the duplicates also we find out the data points which needs to be collected like Insurance Cost , their location of incident , How often they are travelling , What is their income etc .

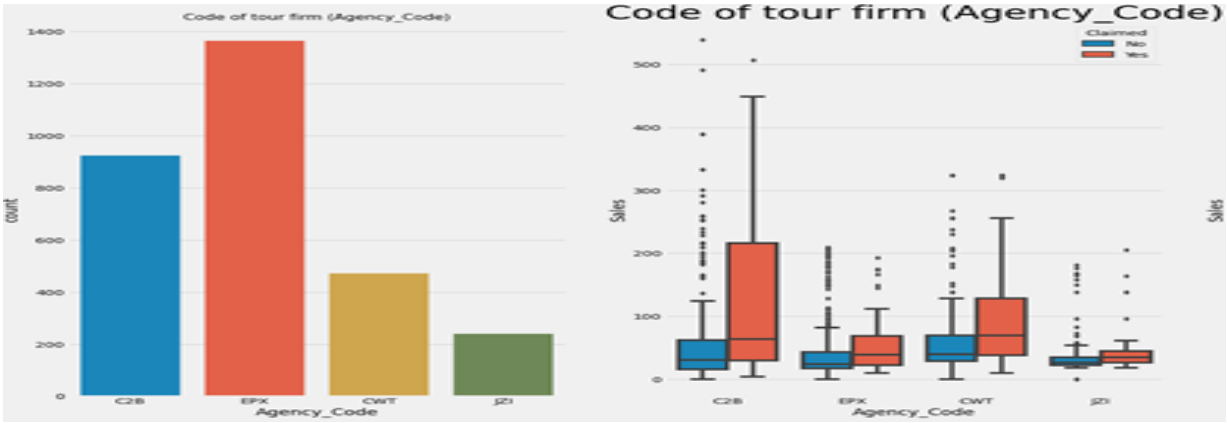
Following are my Suggestions to Management

• Since data set suggest that 90% of insurance done by online mode so we can i ncrease their online experiences and also show the reviews about their insuran ce for benefitted customers, this will increase the Confidence of customer whic h will increase the insurance sales and profit too.

• we have found another interesting fact that more sales happen via Agency th an Airlines and the trend shows the claim are processed more at Airline (See Bo x plot below. So have to find out why this is happening?



• We need to talk with JZI & CWT agency to pick up sales as they Have lower sa les (Refer Below figure) , We can also tell them to use Digital marketing tools to increase their reach to the customers .



• Insurance Company can also increase their Insurance Portfolio by adding mor e insurance plans like health insurance , Corporate accidental Insurance etc to i ncrease their revenue

Below are the some Key performance indicators (KPI) of insurance claims are:

- Average Cost Per Claim
- Claim Frequency
- Average Time to Settle a Claim
- Claims Ratio
- Increase customer satisfaction

Based on the above Insurance Company can make automated reporting proces s to senior officials of the company if a claim is pending from more than 1 week so that genuine claims can't be neglected; this will Build the reputation of the c ompany in the customer's eye.