TRAVEL INSURANCE ANALYSIS

Name Bhushan Rai
PGP-DSBA Online
January' 21
Date: 27/06/2021

Table of Contents

Contents

List of Figures

List of Tables

Executive Summary

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.

Introduction

The purpose of the project is to predict the claim status and provide recommendations based on the best model

Data Description

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)

10Age of insured (Age)

Sample of the dataset:

	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

Table 1. Dataset Sample

Exploratory Data Analysis

Let us check the types of variables in the data frame.

```
---  ------         --------------  -----
 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
```

The dataset consists of 3000 observations where datatypes are 2 float, 2 int, 6 objects

Check for missing values in the dataset:

The dataset has 3000 rows and 10 columns

```
df.isnull().sum()
```

```
Age             0
Agency_Code     0
Type            0
Claimed         0
Commision       0
Channel         0
Duration        0
Sales           0
Product Name    0
Destination     0
dtype: int64
```

There are no missing values in the dataset

Descriptive Statistics:

	count	mean	std	min	25%	50%	75%	max
Age	3000.0	38.091000	10.463518	8.0	32.0	36.00	42.000	84.00
Commision	3000.0	14.529203	25.481455	0.0	0.0	4.63	17.235	210.21
Duration	3000.0	70.001333	134.053313	-1.0	11.0	26.50	63.000	4580.00
Sales	3000.0	60.249913	70.733954	0.0	20.0	33.00	69.000	539.00

The observed values in duration is negative. The mean and median for commission and sales varies significantly

Duplicate Values:

```
Number of duplicate rows = 139
```

	Age	Agency_Code	Type	Claimed	Commision	Channel	Duration	Sales	Product Name	Destination
63	30	C2B	Airlines	Yes	15.0	Online	27	60.0	Bronze Plan	ASIA
329	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
407	36	EPX	Travel Agency	No	0.0	Online	11	19.0	Cancellation Plan	ASIA
411	35	EPX	Travel Agency	No	0.0	Online	2	20.0	Customised Plan	ASIA
422	36	EPX	Travel Agency	No	0.0	Online	5	20.0	Customised Plan	ASIA
...	...	...	...	...	...	...	...	...	...	...
2940	36	EPX	Travel Agency	No	0.0	Online	8	10.0	Cancellation Plan	ASIA
2947	36	EPX	Travel Agency	No	0.0	Online	10	28.0	Customised Plan	ASIA
2952	36	EPX	Travel Agency	No	0.0	Online	2	10.0	Cancellation Plan	ASIA
2962	36	EPX	Travel Agency	No	0.0	Online	4	20.0	Customised Plan	ASIA
2984	36	EPX	Travel Agency	No	0.0	Online	1	20.0	Customised Plan	ASIA

Removed Duplicate values:

```
The duplicated varibales are now removed from the dataset, we have now 2861 rows and 10 columns
```

Univariate Analysis:

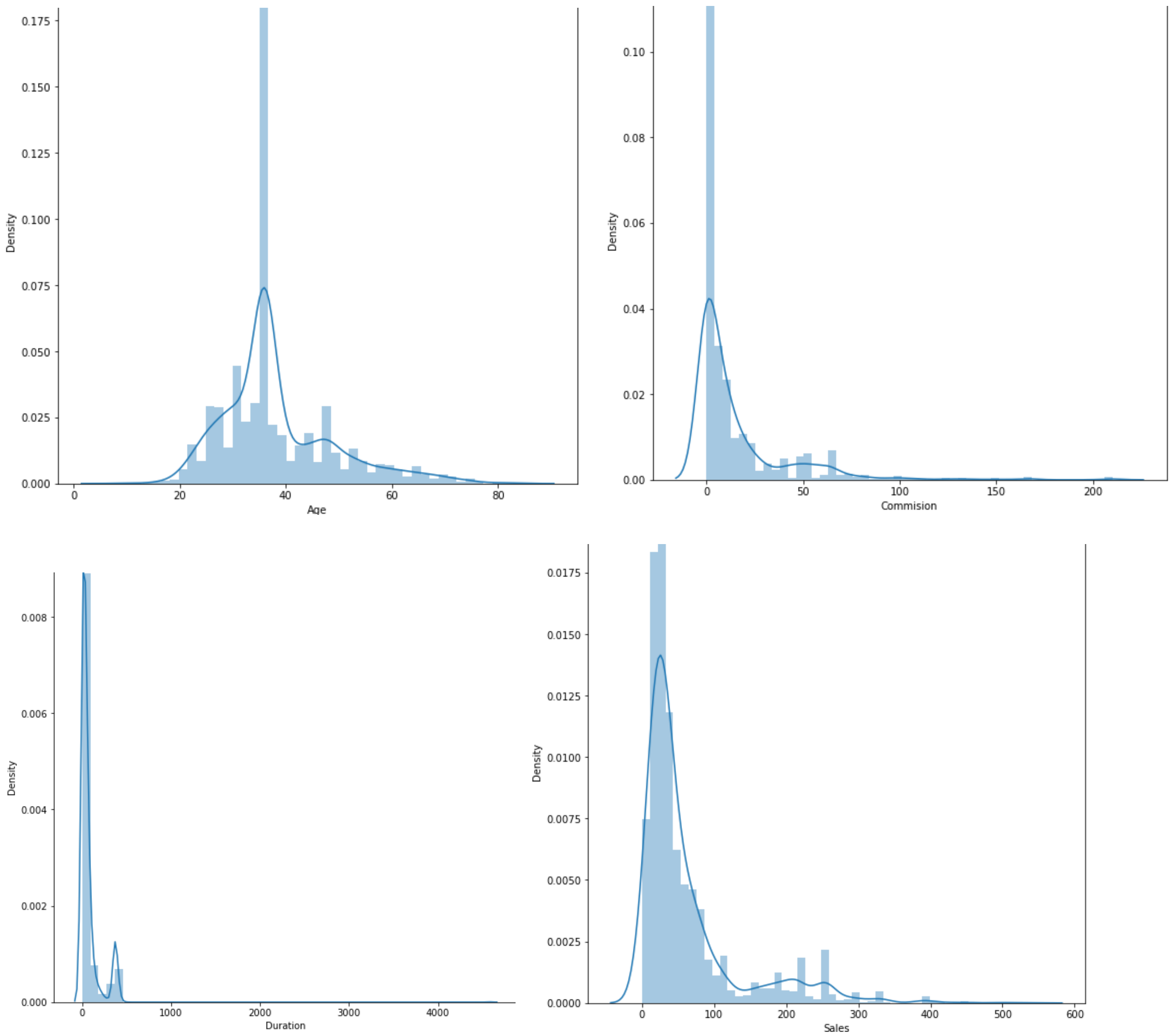
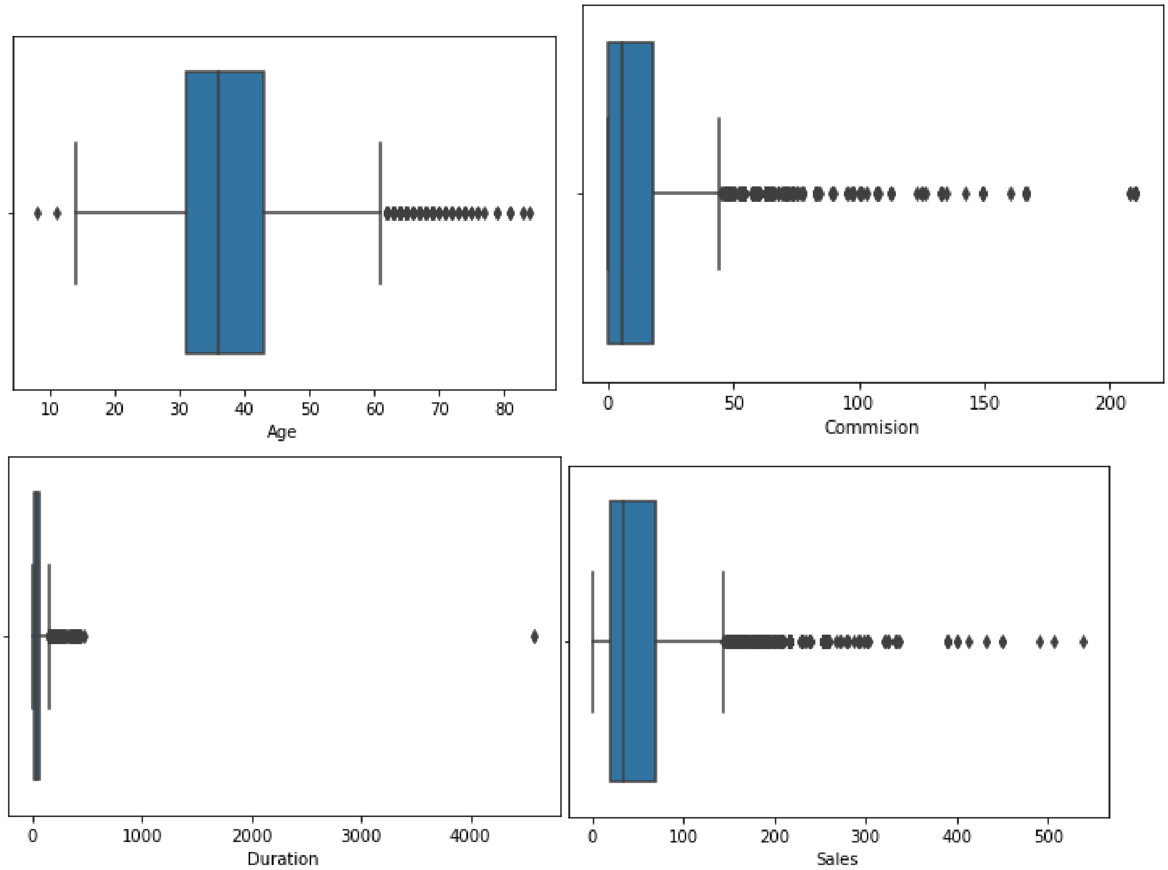


Fig.2 – Distplot

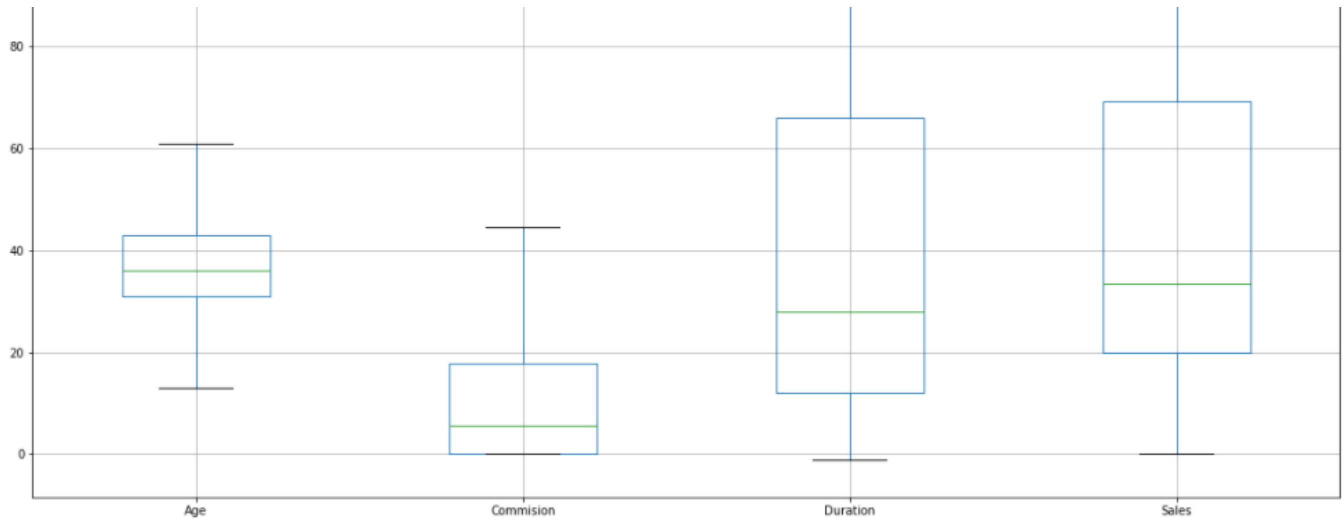
Check for Outliers:



Skewness:

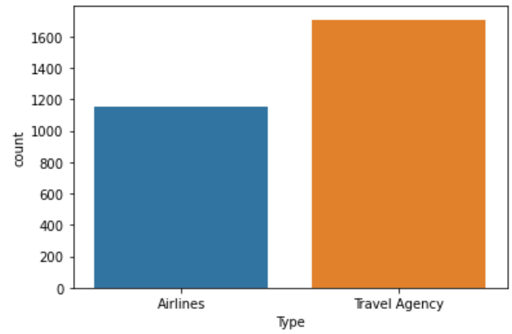
```
Duration      13.786096
Commision      3.104741
Sales          2.344643
Age            1.103145
dtype: float64
```

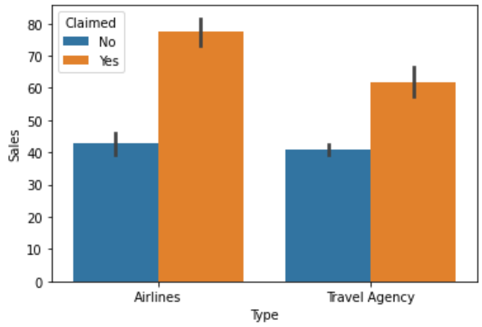
```
As we see there are outliers present in all 4 varibales lets treat them
```

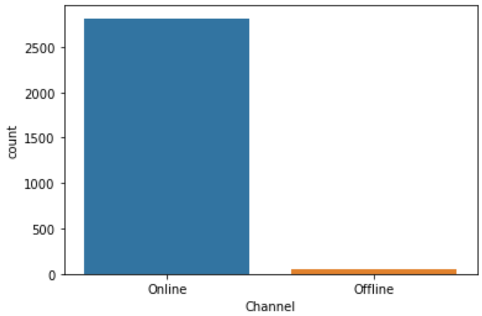


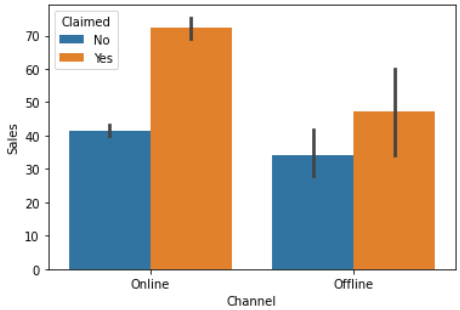
```
The outliers is been treated
```

Count plot:

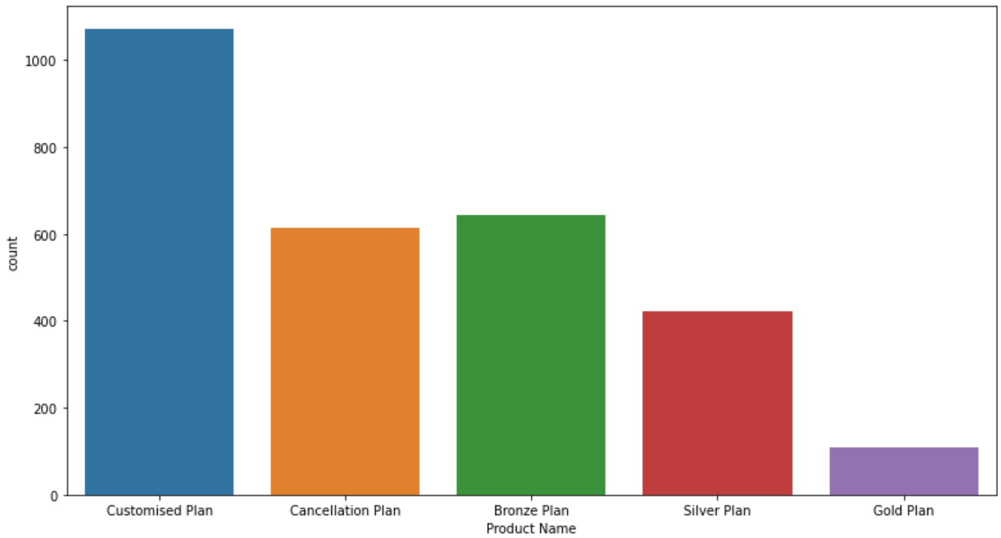


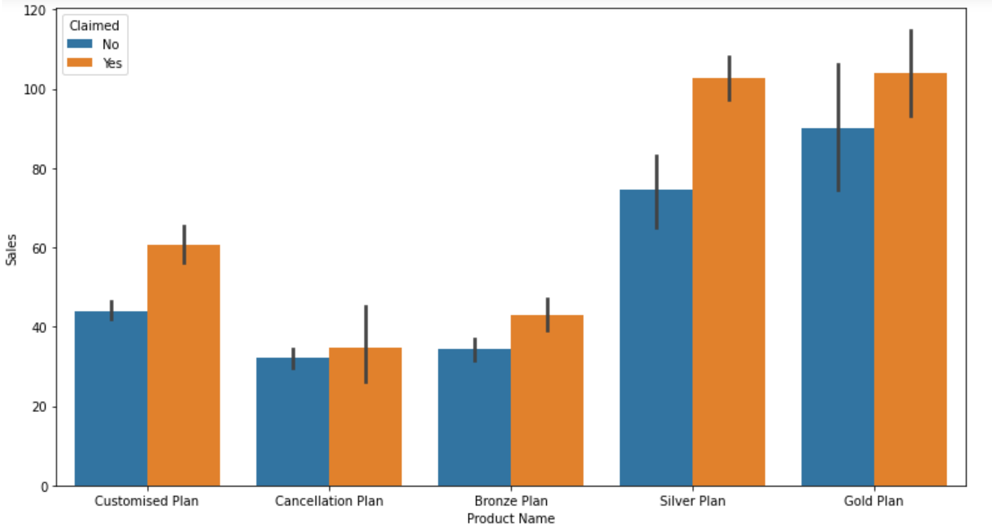


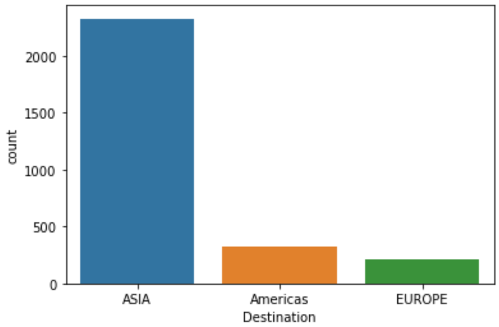


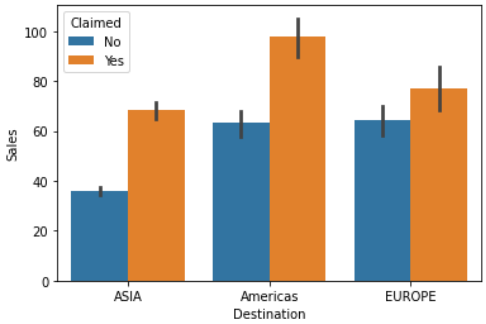


Categorical variables:



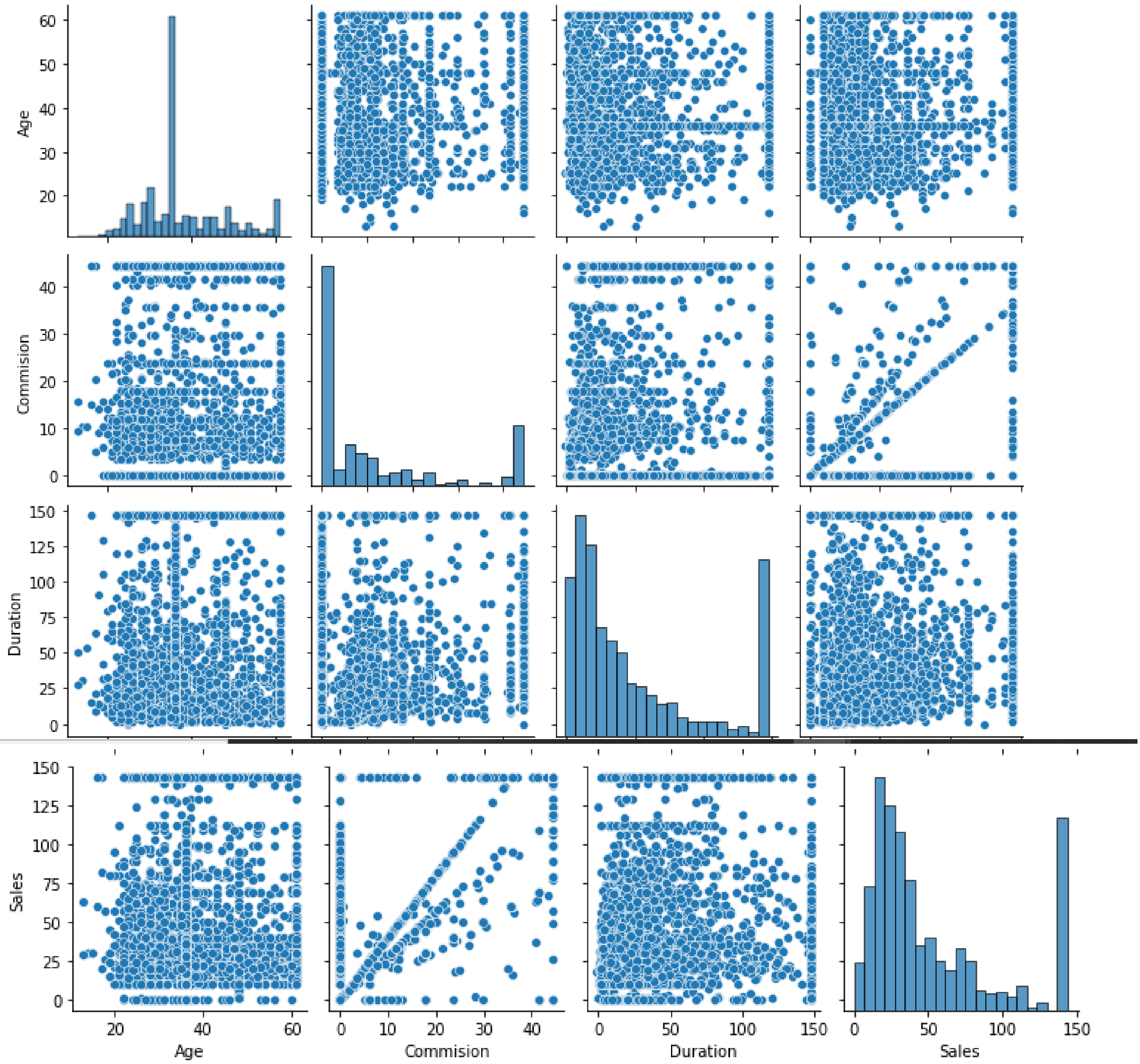






Bivariate Analysis:

Check distribution for continuous distribution:



Check Correlation:

	Age	Commision	Duration	Sales
Age	1.000000	0.071246	0.009216	0.021450
Commision	0.071246	1.000000	0.453225	0.682537
Duration	0.009216	0.453225	1.000000	0.534512
Sales	0.021450	0.682537	0.534512	1.000000

Convert Categorical :

```
 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: 128.5 KB
```

Proportions of 1s and 0s

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

2.2 Data Split: Split the data into test and train(1 pts), build classification model CART (1.5 pts), Random Forest (1.5 pts), Artificial Neural Network(1.5 pts). Object data should be converted into categorical/numerical data to fit in the models. (pd.categorical().codes(), pd.get_dummies(drop_first=True)) Data split, ratio defined for the split, train-test split should be discussed. Any reasonable split is acceptable. Use of random state is mandatory. Successful implementation of each model. Logical reason behind the selection of different values for the parameters involved in each model. Apply grid search for each model and make models on best_params. Feature importance for each model.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy (1 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC_AUC score for each model (2 pts), Make classification reports for each model. Write inferences on each model (2 pts). Calculate Train and Test Accuracies for each model. Comment on the validness of models (overfitting or underfitting) Build confusion matrix for each model. Comment on the positive class in hand. Must clearly show obs/pred in row/col Plot roc_curve for each model. Calculate roc_auc_score for each model. Comment on the above calculated scores and plots. Build classification reports for each model. Comment on f1 score, precision and recall, which one is important here.

Cart Model:

Dimensions of train and test data:

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

Variable Importance:

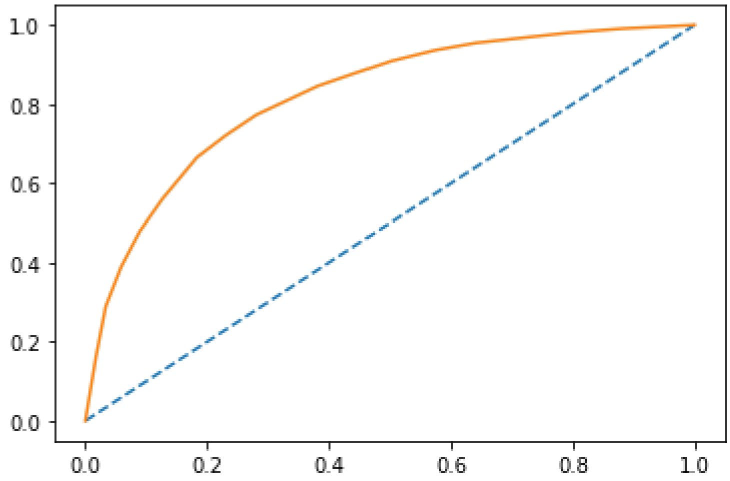
	Imp
Age	0.174976
Agency_Code	0.204343
Type	0.001882
Commision	0.079623
Channel	0.002774
Duration	0.223499
Sales	0.230417
Product Name	0.059610
Destination	0.022875

Predicted Class and Probs:

	0	1
0	0.842105	0.157895
1	0.923077	0.076923
2	0.480392	0.519608
3	0.633663	0.366337
4	0.842105	0.157895

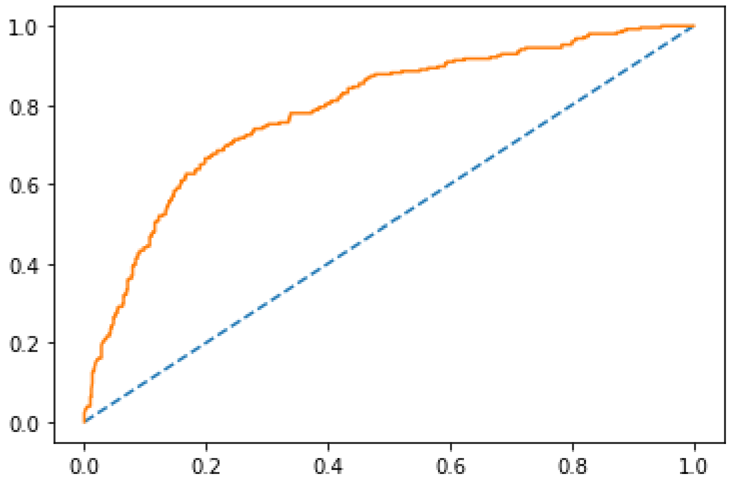
AUC for training Data

AUC: 0.820



AUC for test Data

AUC: 0.789



Train Data:

	precision	recall	f1-score	support
0	0.80	0.87	0.84	1342
1	0.69	0.56	0.62	660
accuracy			0.77	2002
macro avg	0.74	0.72	0.73	2002
weighted avg	0.76	0.77	0.76	2002

Test Data:

	precision	recall	f1-score	support
0	0.81	0.85	0.83	605
1	0.60	0.52	0.56	254
accuracy			0.75	859
macro avg	0.70	0.69	0.69	859
weighted avg	0.75	0.75	0.75	859

```
cart_test_precision  0.6
cart_test_recall  0.52
cart_test_f1  0.56
```

Cart Conclusion: cart_train_precision 0.69 cart_train_recall 0.56 cart_train_f1 0.62

cart_test_precision 0.6 cart_test_recall 0.52 cart_test_f1 0.56

The train and test data are almost similar, the model seems okay.

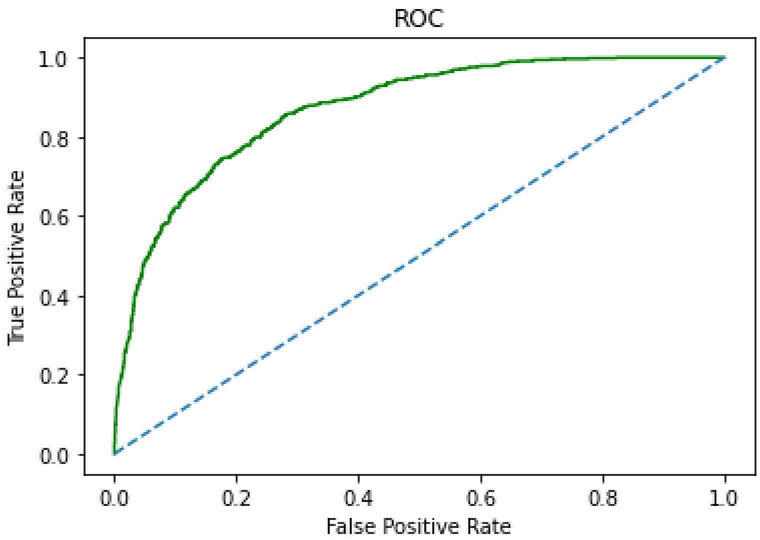
Random Forest:

Train Data:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	1342
1	0.74	0.62	0.68	660
accuracy			0.80	2002
macro avg	0.78	0.76	0.77	2002
weighted avg	0.80	0.80	0.80	2002

```
rf_train_precision  0.74
rf_train_recall  0.62
rf_train_f1  0.68
```

```
Area under Curve is 0.8707638982974303
```

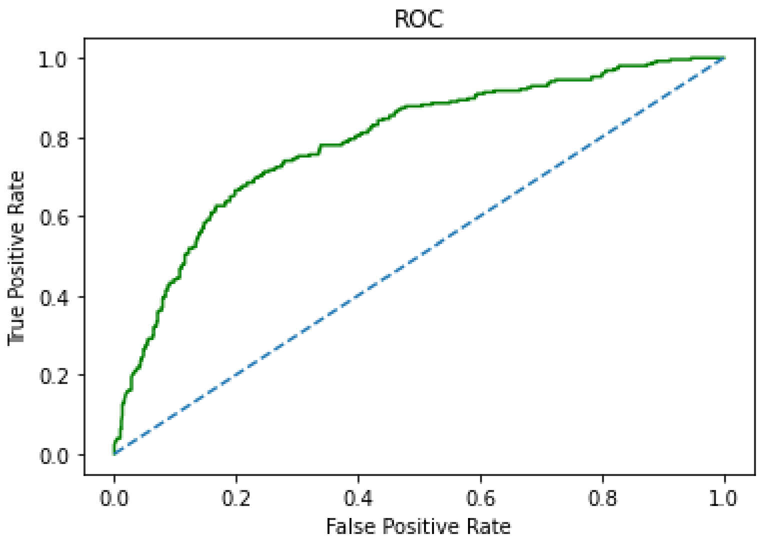


Test Data:

	precision	recall	f1-score	support
0	0.82	0.86	0.84	605
1	0.62	0.56	0.59	254
accuracy			0.77	859
macro avg	0.72	0.71	0.71	859
weighted avg	0.76	0.77	0.76	859

```
rf_test_precision  0.62
rf_test_recall  0.56
rf_test_f1  0.59
```

Area under Curve is 0.7885078414784928



	Imp
Agency_Code	0.378198
Sales	0.193812
Product Name	0.182063
Duration	0.089622
Age	0.067383
Commision	0.061467
Type	0.018426
Destination	0.008378
Channel	0.000651

Random Forest Conclusion: rf_test_precision 0.62 rf_test_recall 0.56 rf_test_f1 0.59

rf_train_precision 0.74 rf_train_recall 0.62 rf_train_f1 0.68

Test seems to be performing better here , could be overfitting, however with overall can be considered as good model

ANN model:

Dimensions of train and test data:

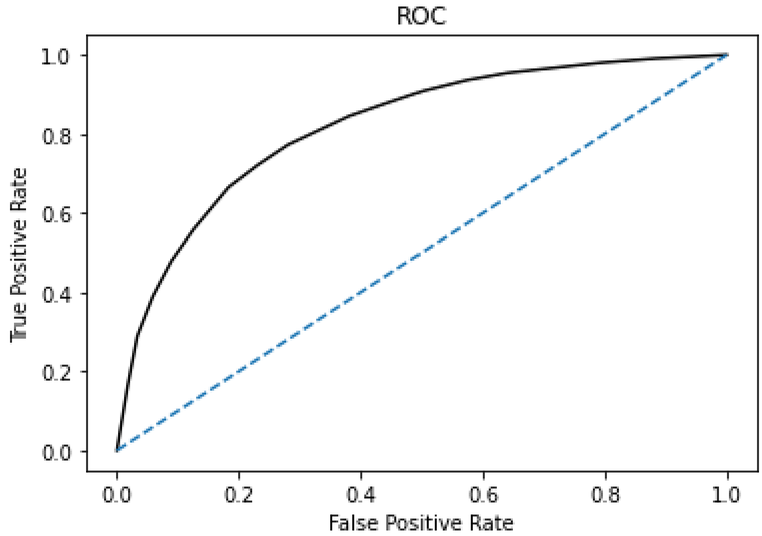
```
X_train:  (2002, 9)
X_test:  (859, 9)
y_train:  (2002,)
y_test:  (859,)
```

Train Data:

	precision	recall	f1-score	support
0	0.80	0.87	0.84	1342
1	0.69	0.56	0.62	660
accuracy			0.77	2002
macro avg	0.74	0.72	0.73	2002
weighted avg	0.76	0.77	0.76	2002

```
nn_train_precision  0.69
nn_train_recall  0.56
nn_train_f1  0.62
```

```
Area under Curve is 0.8203862394436165
```

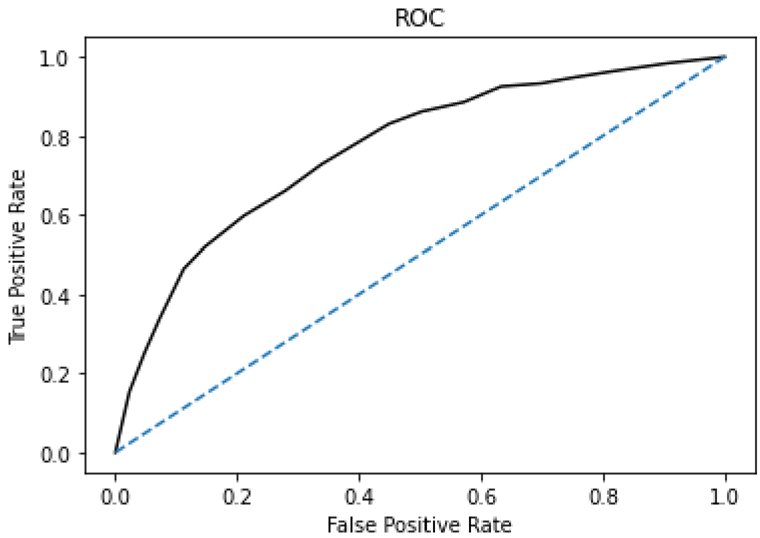


Test Data:

	precision	recall	f1-score	support
0	0.81	0.85	0.83	605
1	0.60	0.52	0.56	254
accuracy			0.75	859
macro avg	0.70	0.69	0.69	859
weighted avg	0.75	0.75	0.75	859

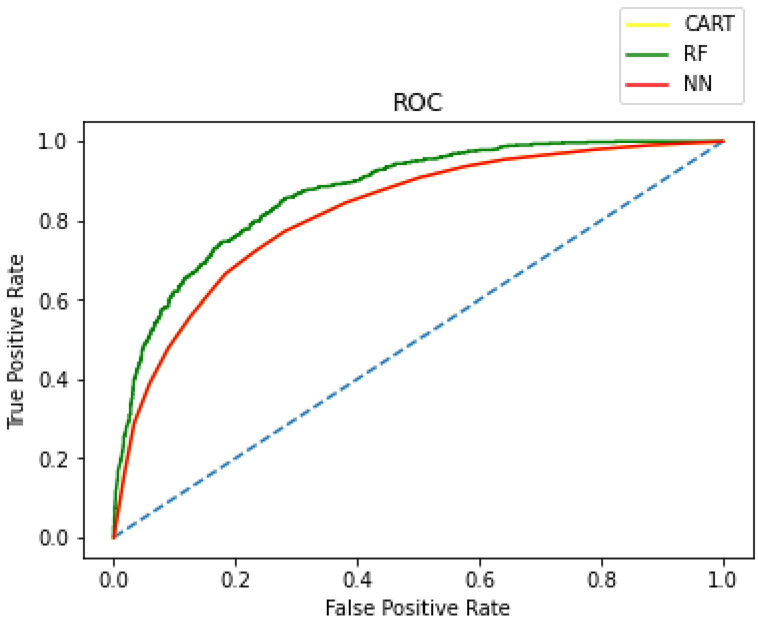
```
nn_test_precision  0.6
nn_test_recall  0.52
nn_test_f1  0.56
```

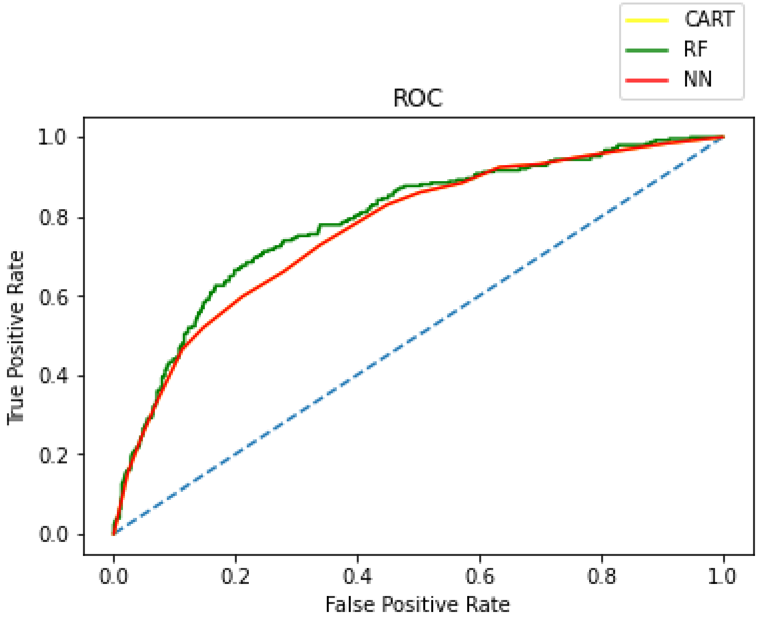
```
Area under Curve is 0.7674952820980022
```



Comparison of performance metrics for 3 models:

	CART Train	CART Test	Random Forest Train	Random Forest Test	Neural Network Train	Neural Network Test
Accuracy	0.77	0.75	0.80	0.77	0.77	0.75
AUC	0.77	0.77	0.87	0.79	0.82	0.77
Recall	0.56	0.52	0.62	0.56	0.56	0.52
Precision	0.69	0.60	0.74	0.62	0.69	0.60
F1 Score	0.62	0.56	0.68	0.59	0.62	0.56





#2.5 Based on your analysis and working on the business problem, detail out appropriate insights and recommendations to help the management solve the business objective. There should be at least 3-4 Recommendations and insights in total. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks should only be allotted if the recommendations are correct and business specific.

Random Forest seems to performing best of the 3 models, with better accuracy, precision, recall and F1 score

As we see the maximum insurance is booked thorugh online channel and very few from offline channel.Customers are benefitting from the source however can be seen offline has claims associated with it. Recommedend to run promotional campigns for other areas so project sales can be boosted. As noticed the claimed is higher on gold plan however cutomized plan shows higher count, as well as for the destination Asia seems to have a higher count however claimed is from other regions.

We would need to collect more data on real time basis.

Recomended:

1. Marketing offers to launch new campaigns
2. Reduce Claim cycle
3. optimize claim recovery
4. Reduce claim handling
5. Increase customer satisfaction

The End