**DATA MINING BUSINESS REPORT**

***Abhishek Chauhan***

***Mrinalini Kulkarni***

***Pavan Krovi***

***Nivedita Kulkarni***

***Yogesh Bhangale***

***PGPDSA DEC’19 BATCH – Group 6***

***Date: Dec 27, 2020***

**CONTENTS**

1. Objective ………………………………………………………………………………………………… 3
2. Problem 1: CART, RF and ANN Model for Insurance Claim.……………………….4 – 21
3. Assumptions …………………………………………………………………………………4
4. Importing Packages ………………………………………………………………………4
5. Solution 1.1 ………………………………………………………………………………….5
6. Solution 1.2 ………………………………………………………………………………….15
7. Solution 1.3 ………………………………………………………………………………….18
8. Solution 1.4 ………………………………………………………………………………….20
9. Solution 1.5 ………………………………………………………………………………….21

**PROBLEM 1: CART-RF-ANN**

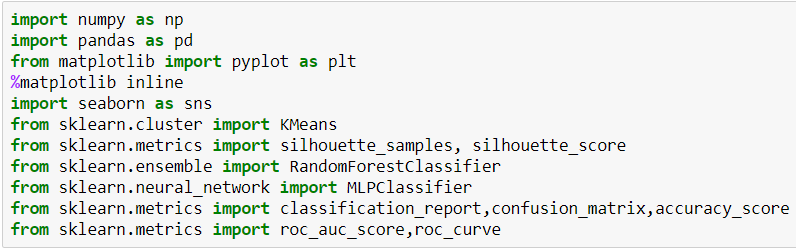
**Assumptions:**

The dataset provided to us is stored as “insurance\_part2\_data.csv” which contains data of 3000 customers and 10 variables namely:

|  |  |
| --- | --- |
| Age | Age of insured |
| Agency\_Code | Code of tour firm |
| Type | Type of tour insurance firms |
| Claimed | Target: Claim Status |
| Commission | The commission received for tour insurance firm |
| Channel | Distribution channel of tour insurance agencies |
| Duration | Duration of the tour |
| Sales | Amount of sales of tour insurance policies |
| Product Name | Name of the tour insurance products |
| Destination | Destination of the tour |

**IMPORTING PACKAGES**

To import the dataset and perform Exploratory Data Analysis on the given dataset we imported the following packages:

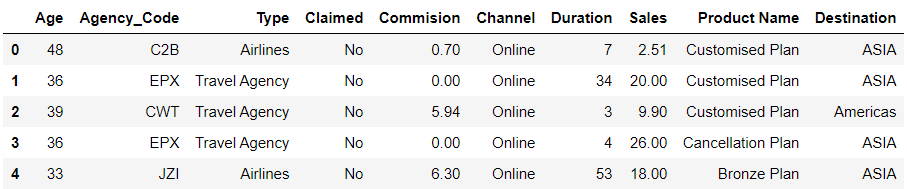


**Solution:**

**1.1 To read the dataset and perform the descriptive statistics and do null value condition check and write an inference on it.**

**Importing the Dataset**

The dataset in question is imported in jupyter notebook using **pd.read\_csv ()** function and will store the dataset in “**claim\_df**”. The top 5 rows of the dataset are viewed using **pd.head ()** function.



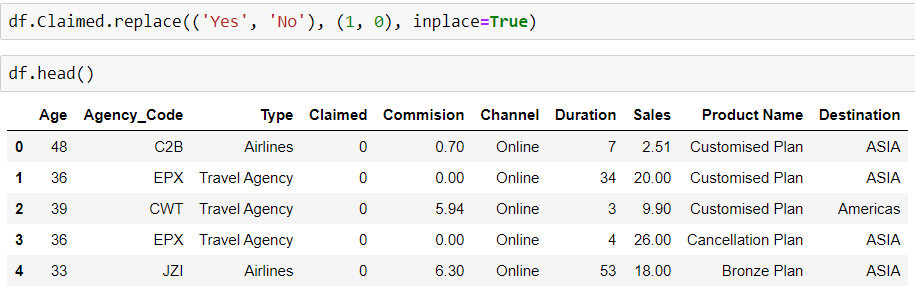
**Observations from Basic EDA:**

-Total Number of variables are 10.

-9 independent variable and one target variable – Claimed.

-Age, Commission, Duration, Sales are numeric variable and remaining are categorial variables

Claimed column Yes and No are replaced with 1 and 0 respectively as below-



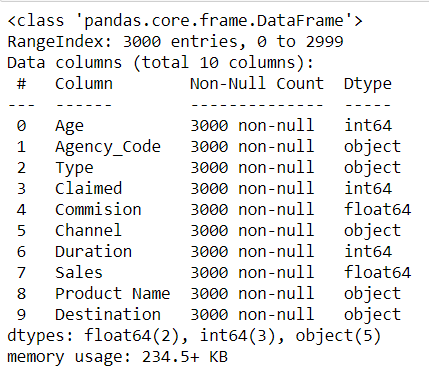
**Dimension of the Dataset**

Let us check the number of rows and the number of columns in the dataframe-



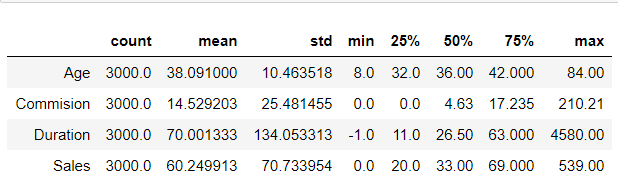
**Structure of the Dataset**

Structure of the Dataset can be computed using **df.info()** function.



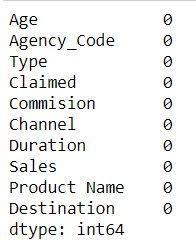
**Summary of the Dataset**

The summary of the dataset can be computed using pd.describe () function.



**Checking for Missing Values**

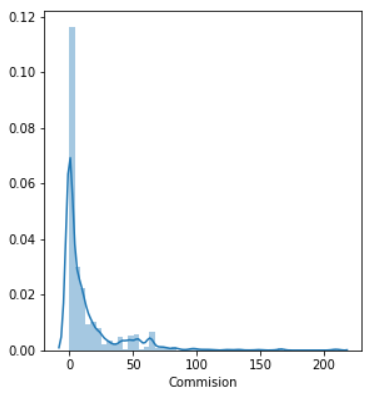
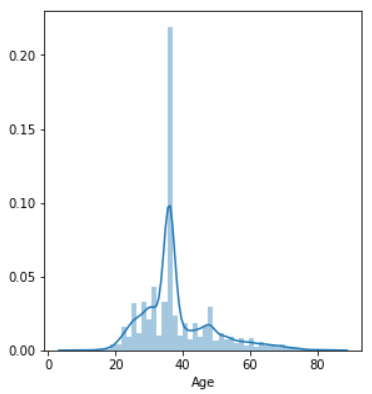
The missing values or “NA” needs to be checked and dropped from the dataset for the ease of evaluation and null values can give errors or disparities in results. Missing Values can be computed using **.isnull().sum()** function.

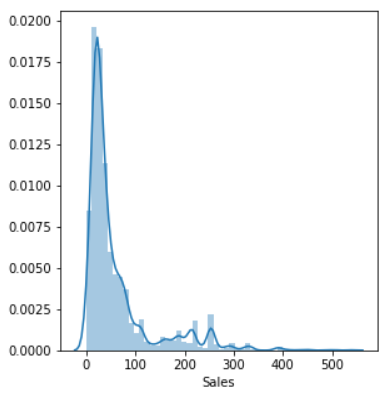
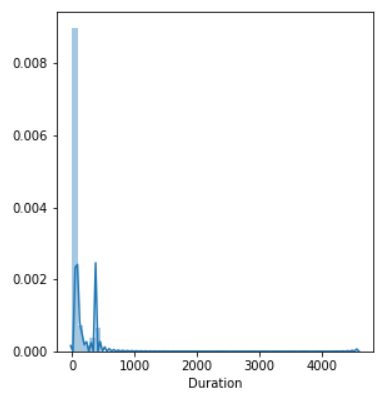


There are no missing values in the dataset hence we can go ahead with building the model.

**Univariate Analysis**

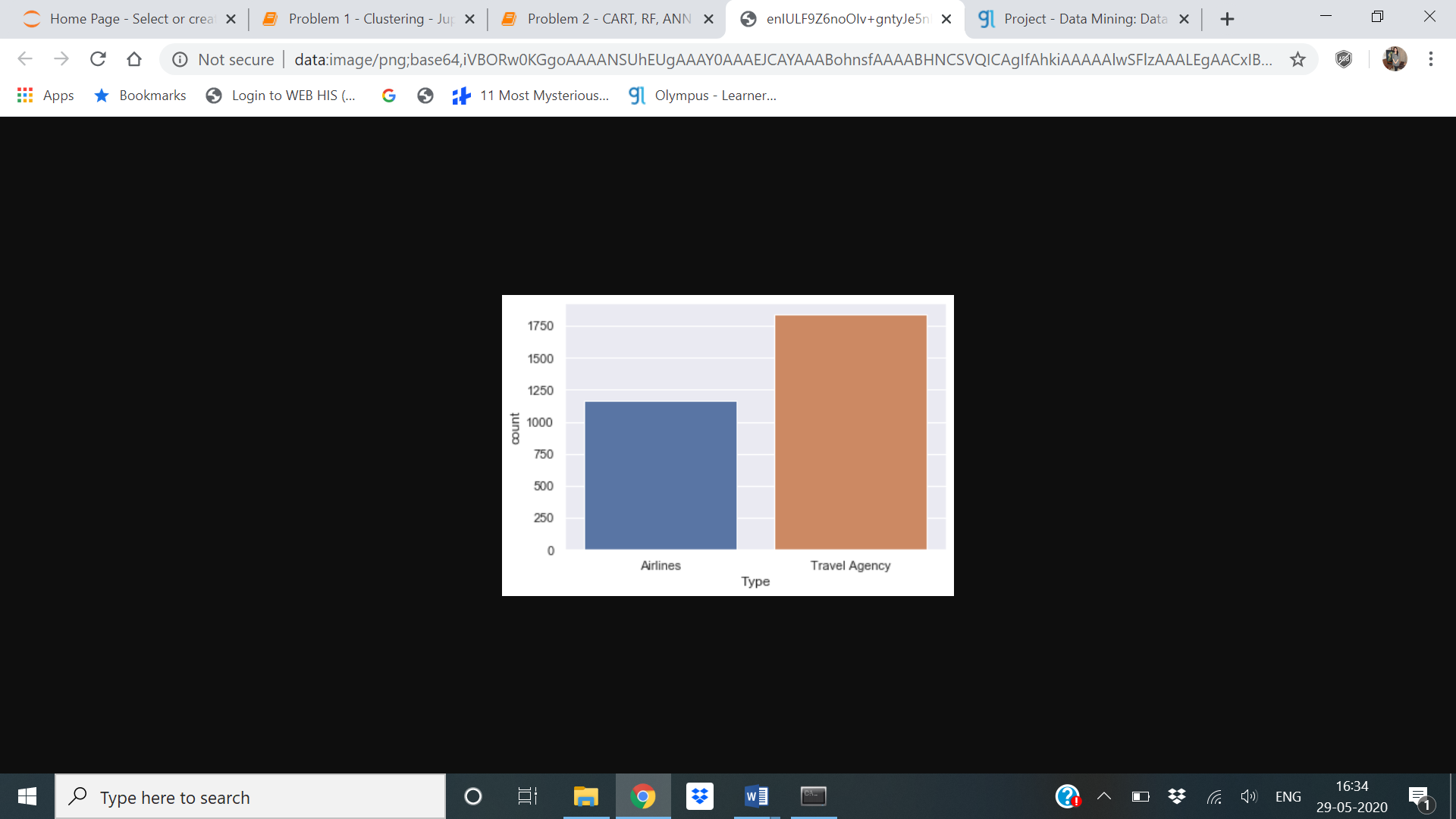
Histograms are plotted for all the numerical variables using **sns.distplot ()** function from seaborn package.

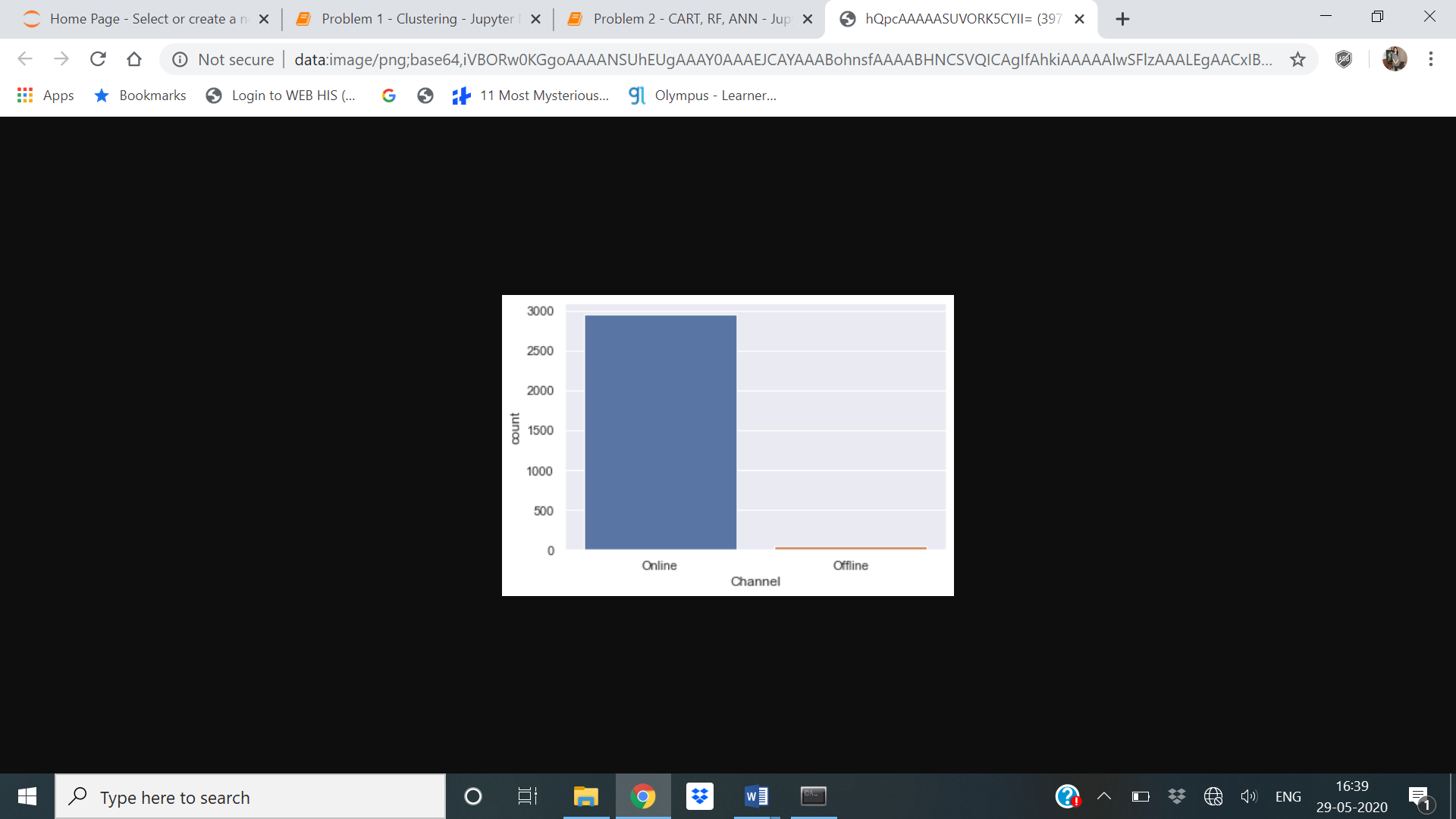




We can see that data is rightly skewed.

Bar Plots are plotted for all Categorical Variables using **sns.countplot()** function from seaborn package.







**Inferences from bar plot-**

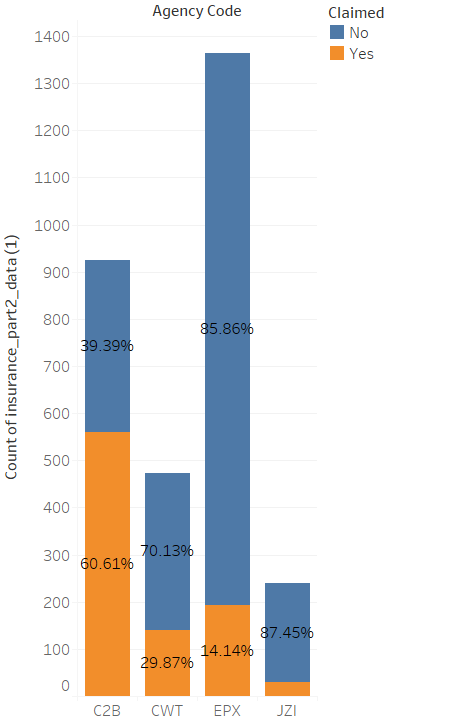
Highest insurance is from Travel Agency.

Highest insurance distribution channel of tour insurance agencies is online.

Highest insurance product plan is customised.

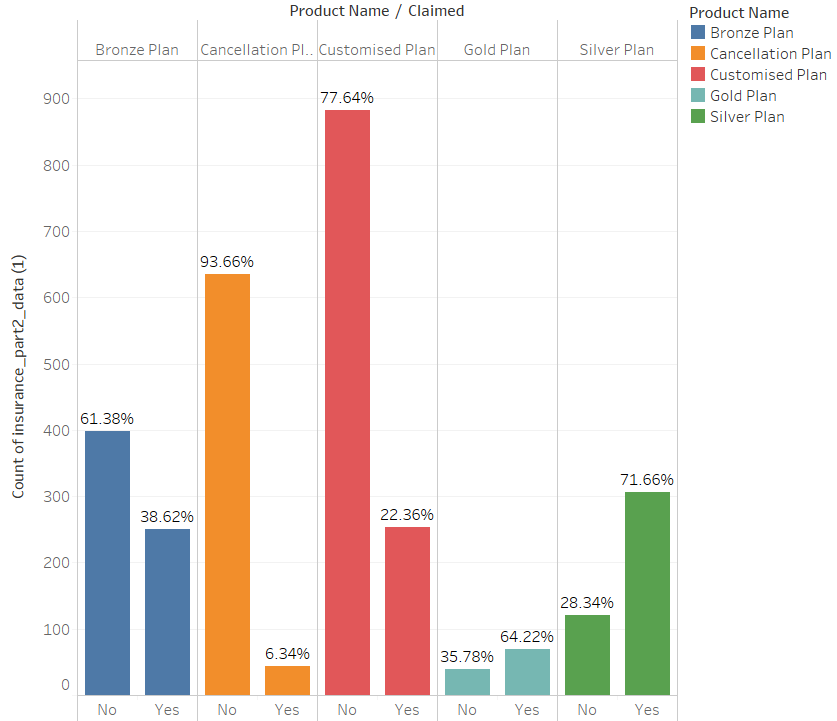
Highest travel destination is Asia.

**Insights from Tableau:**

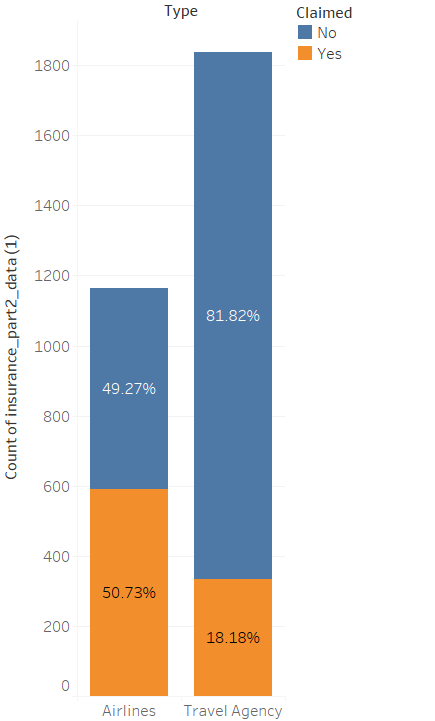


Maximum claims have been raised for the C2B Agency.

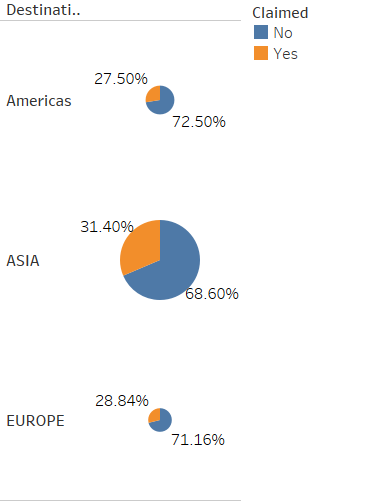
This probably indicates that a closer look needs to be taken at the client list for the C2B Agency.



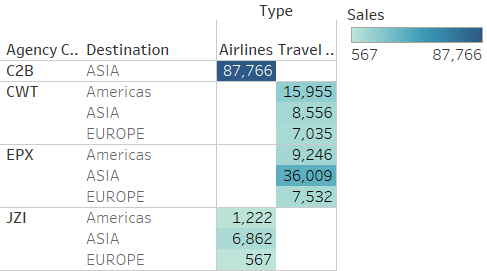
The highest claims are for the Silver Plan followed by the Gold plan. The lowest claims are for the Cancellation Plan. This would probably indicate that the Cancellation plan features can be adopted / customized for the Gold and Silver plans too.



The claims for Airlines are roughly 3 times that of the Travel Agency. This would probably mean that a further analysis for Airlines should be carried out to understand why the claim rate is significantly higher.

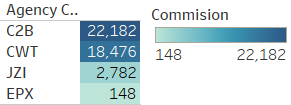


ASIA accounts for the highest number of sales and also the highest number of claims.



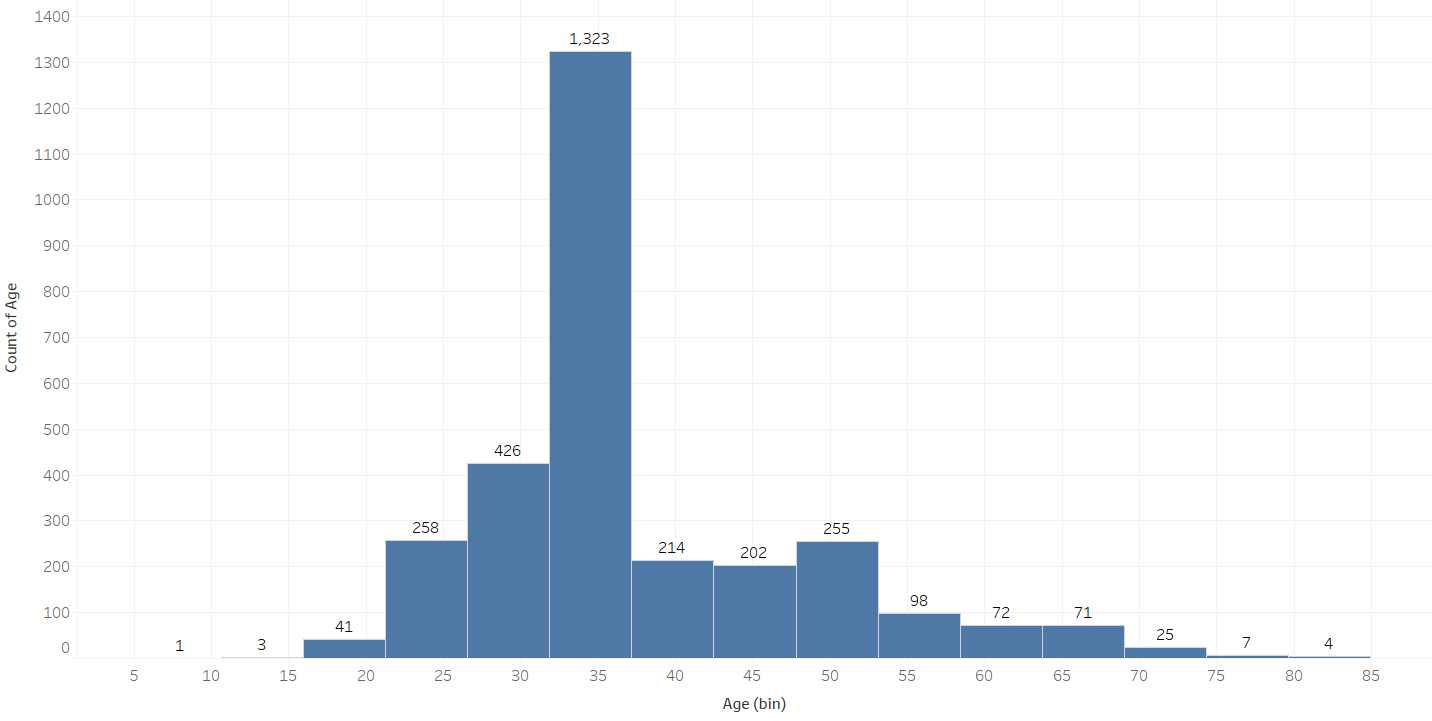
The C2B agency operates only in ASIA. The C2B Agency has the highest number of sales.

Sales across ASIA are highest for all agencies while EUROPE constitutes for the lowest number of sales. This probably indicates that better offers and increased marketing can be directed for Europe and the Americas.



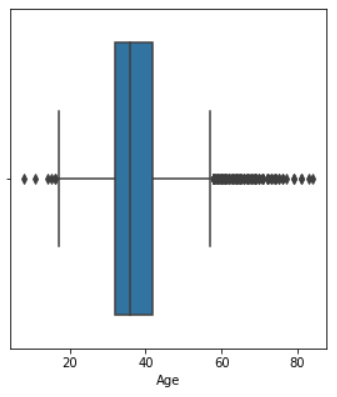
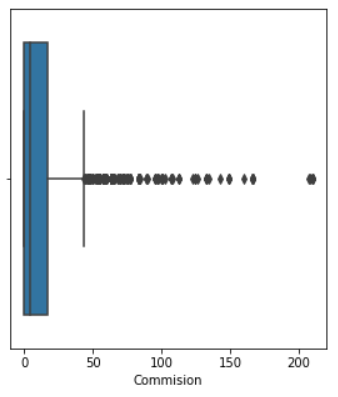
C2B agency earns the highest commissions followed by CWT. EPX is doing the least business.

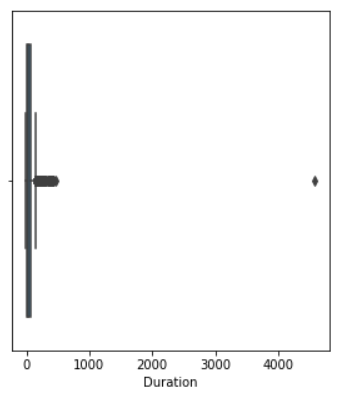
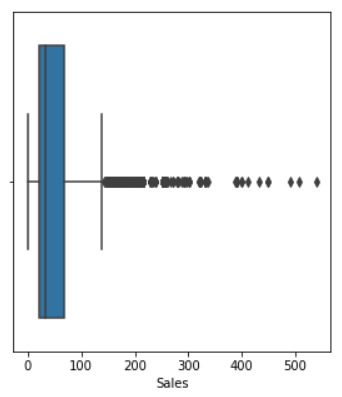
This probably indicates that better training needed for EPX employees.



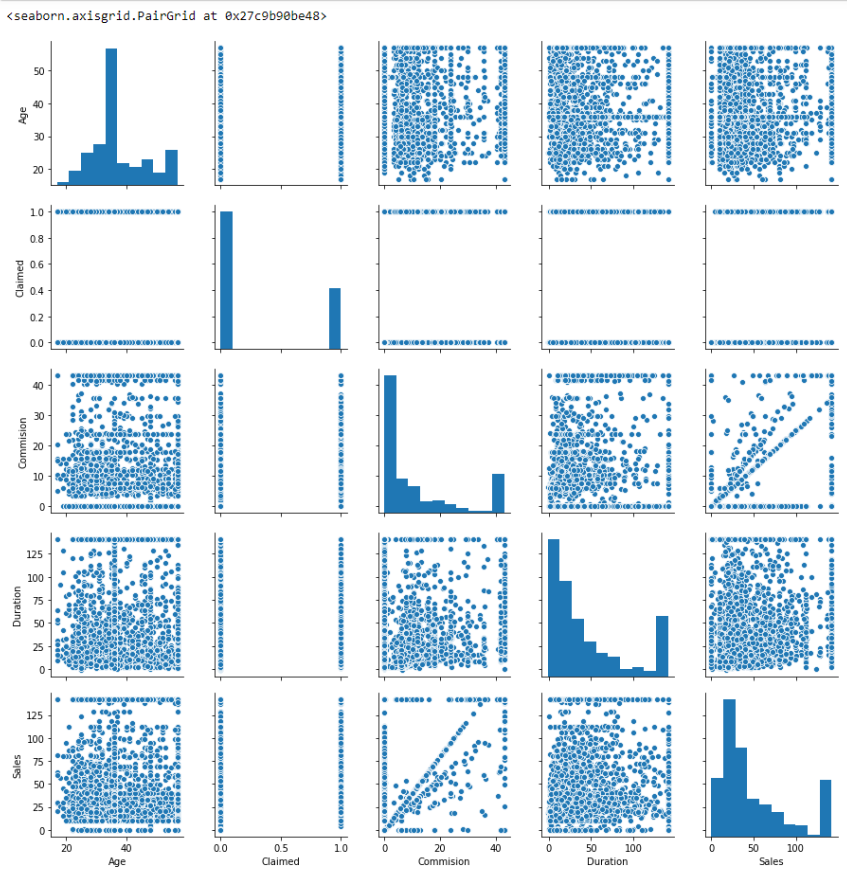
The maximum customers for this travel agency are in the age bracket of 30-35years. This could probably mean that there is scope to increase customers in older age brackets by introducing better packages, schemes and plans for the senior citizens.

**Boxplots of Variables to check for Outliers**

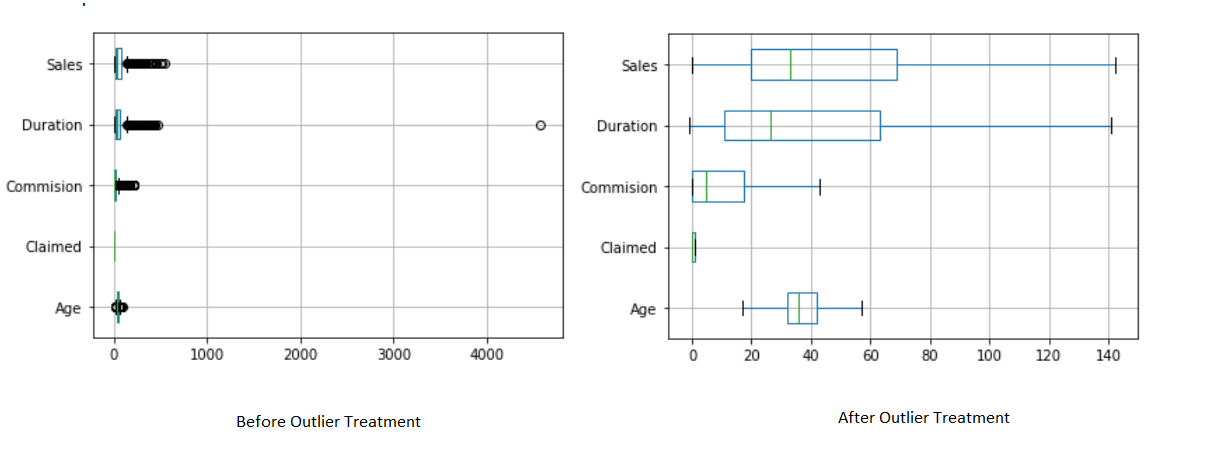
 

**Bivariate**



**Outliers Treatment**

Combined boxplot for all variables before outlier treatment–

****

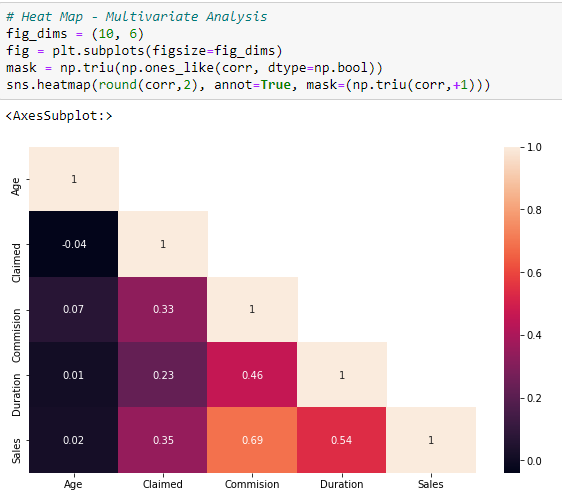
**Inference**: After plotting the Boxplots for all the numerical variables we can conclude that a very high number of outliers are present in the variables namely, **Age, Commission, Duration and Sales** which means that we need to treat these outlier values so as to proceed further with our model building and analysis as these values can create errors and can deviate from the actual results.

We can conclude from the above graphs that the majority of the customers doing a claim in our data belong to age group of 25-40 with the type of Tour Agency firm being Travel Agency, Channel being Online, Product name being Customised Plan and Destination being Asia.

**Multivariate Analysis**

**Heat Map (Relationship Analysis)**

We will now plot a Heat Map or Correlation Matrix to evaluate the relationship between different variables in our dataset. This graph can help us to check for any correlations between different variables.



Interpretation from the above heat map – There is highest corelation between Commission and sales.

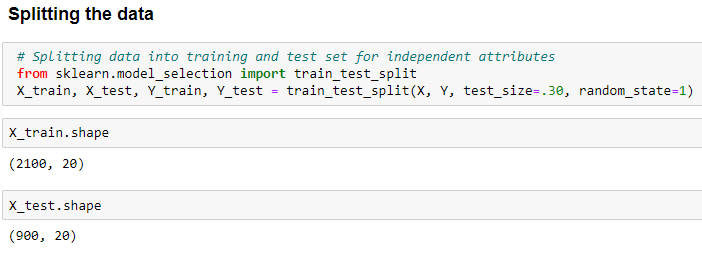
**1.2. To split the data into test and train, build classification model CART, Random Forest and Artificial Neural Network.**

**Converting ‘object’ datatype to ‘int’**

For our analysis and building Decision tree and Random Forest, we have to convert the variables which have ‘object’ datatype and convert them into integer. Hence we used pd.get\_dummies to convert categorical data into dummy variables.

**Splitting Dataset in Train and Test Data (70:30)**

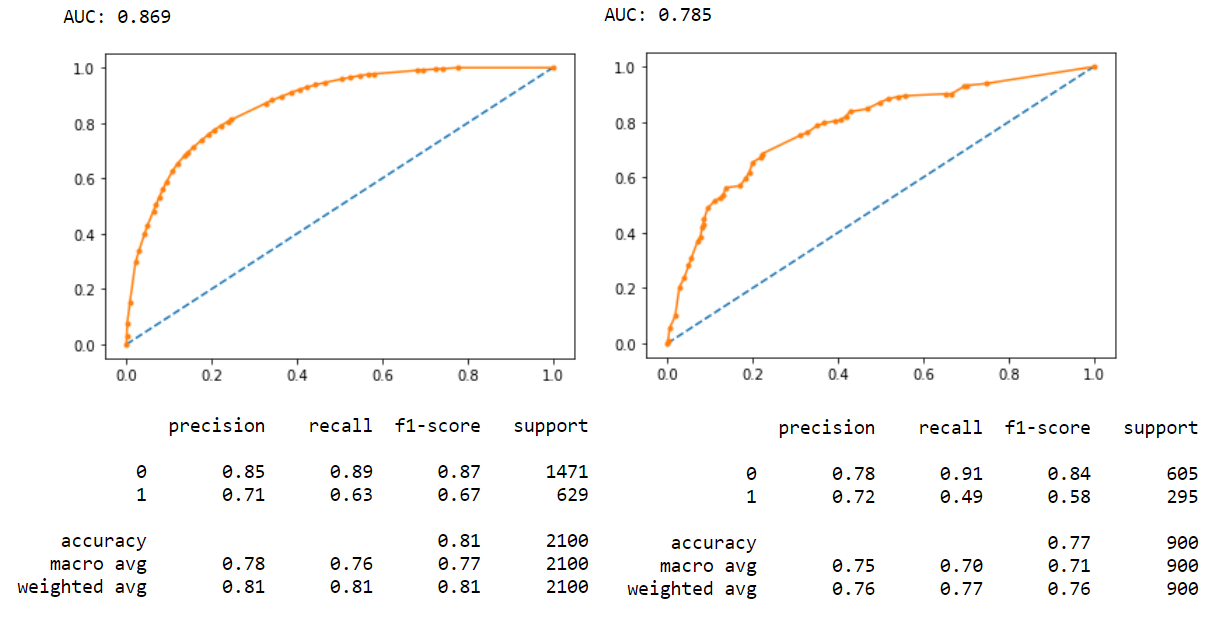
For building the models we will now have to split the dataset into Training and Testing Data with the ratio of 70:30. These two datasets are stored in X\_train and X\_test with their corresponding dimensions as follows-



**1.3 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**

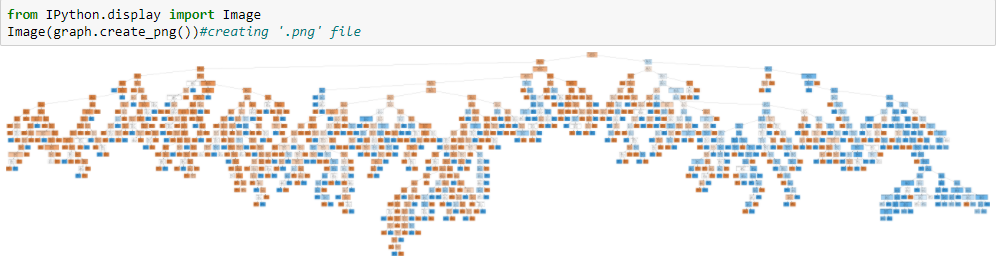
### 1] Decision Tree - CART

|  |  |
| --- | --- |
| Train | Test |
| True Negative: 1312 | True Negative: 548 |
| False Positives: 159 | False Positives: 57 |
| False Negatives: 235 | False Negatives: 151 |
| True Positives: 394 | False Negatives: 144 |

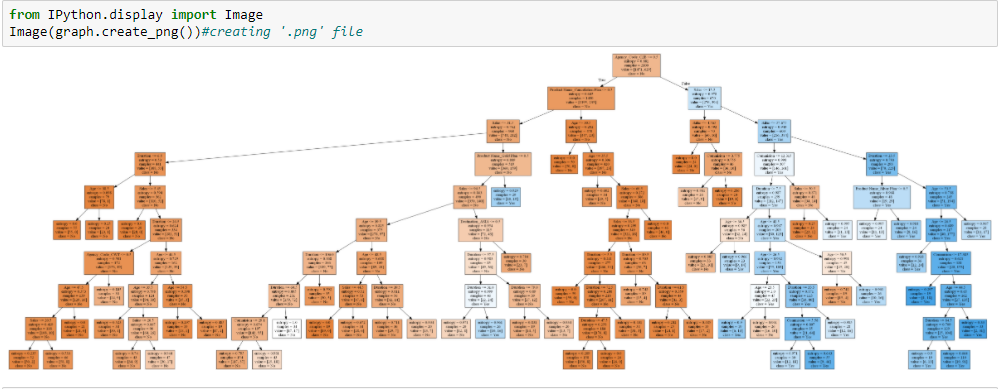
****

We have been able to predict 71% of the target variables correctly.

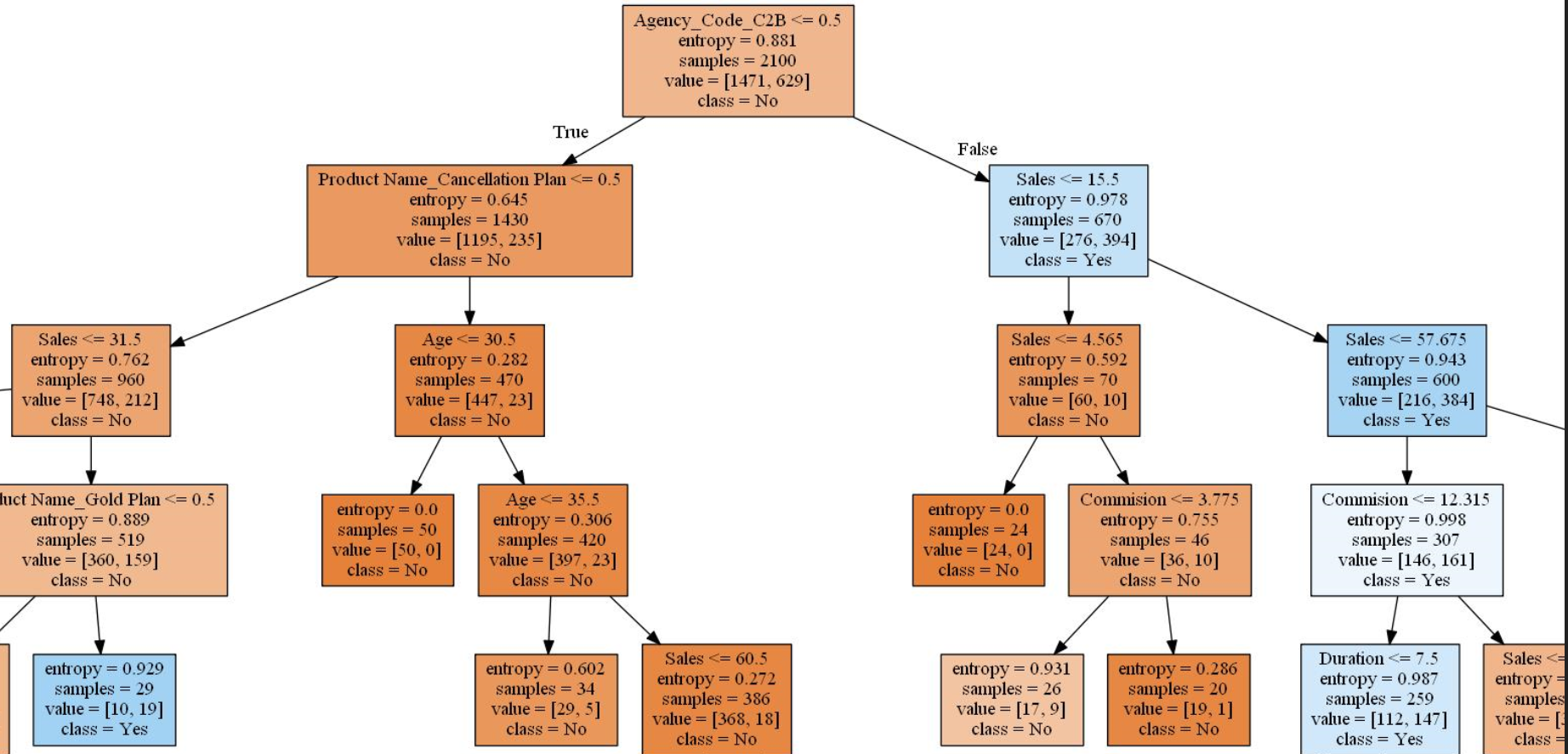
**Before Pruning-**



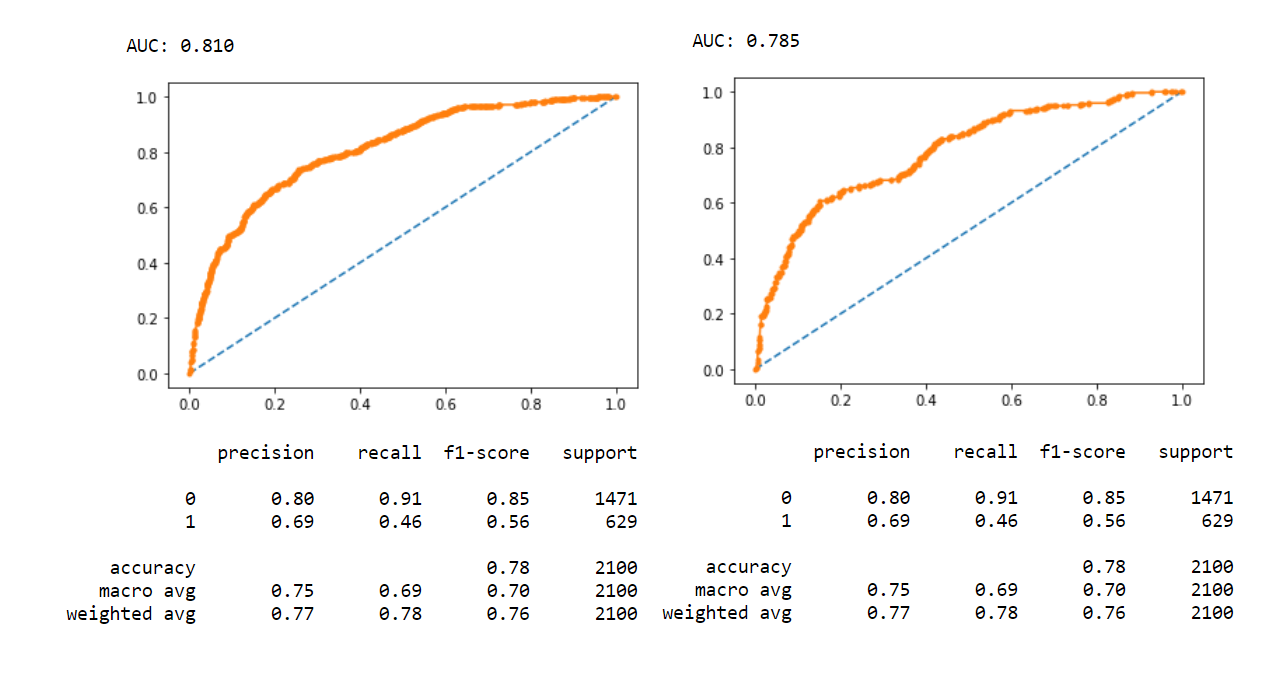
**After Pruning-**



After zooming in higher weightage can be seen for agency code.



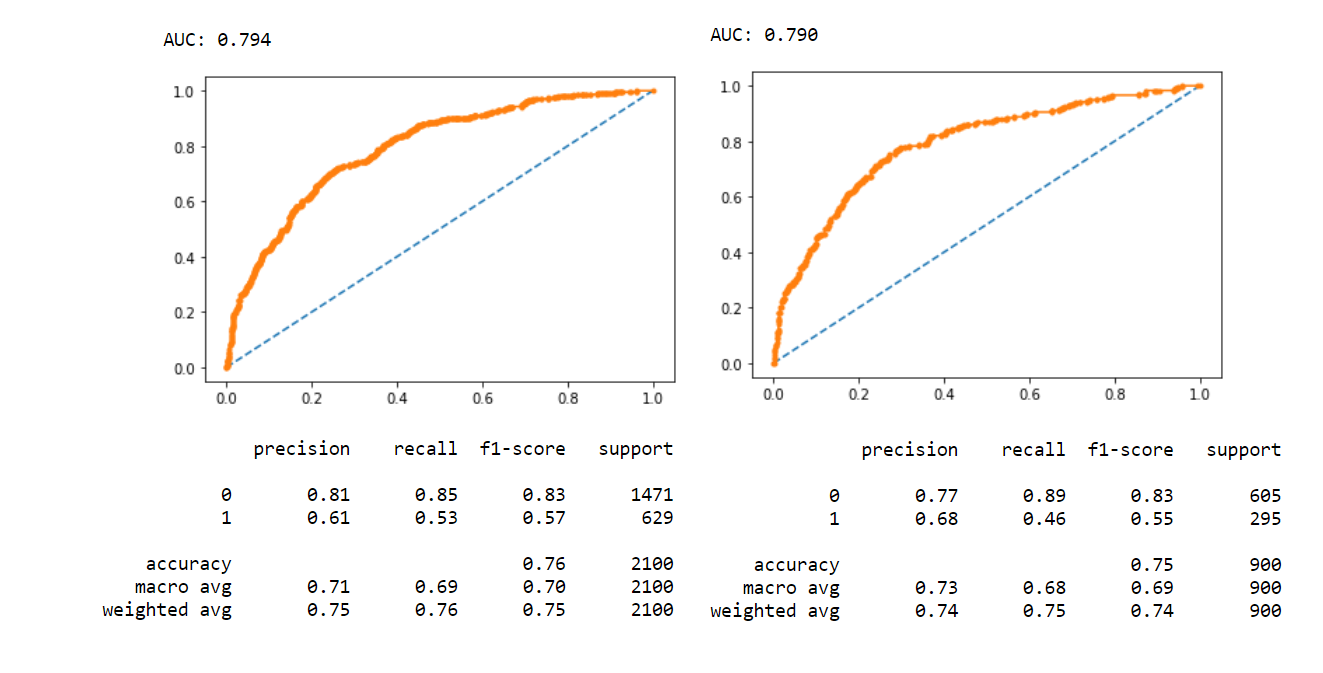
**2] Random Forest**



We have been able to predict 69% of the target variables correctly.

|  |  |
| --- | --- |
| Train | Test |
| True Negative: 1343 | True Negative: 567 |
| False Positives: 128 | False Positives: 38 |
| False Negatives: 348 | False Negatives: 186 |
| True Positives: 291 | False Negatives: 109 |

3] ANN :



We have been able to predict 61% of the target variables correctly.

|  |  |
| --- | --- |
| Train | Test |
| True Negative: 1255 | True Negative: 541 |
| False Positives: 216 | False Positives: 64 |
| False Negatives: 294 | False Negatives: 160 |
| True Positives: 335 | False Negatives: 135 |

On comparing these three models Random forest looks suitable models where the accuracy is 0.78.

**1.4. To compare all the models and write an inference which model is best/optimized.**

Comparison of all the performance evaluators for the three models are given in the following table. We are using Precision, F1 Score and AUC Score for our evaluation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Precision** | **Accuracy** | **AUC Score** |
| **CART Model** |  |  |  |
| *Train Data* | 0.71 | 0.81 | 0.86 |
| *Test Data* | 0.72 | 0.71 | 0.78 |
| **Random Forest** |  |  |  |
| *Train Data* | 0.69 | 0.78 | 0.81 |
| *Test Data* | 0.69 | 0.78 | 0.78 |
| **Neural Network** |  |  |  |
| *Train Data* | 0.61 | 0.76 | 0.79 |
| *Test Data* | 0.68 | 0.75 | 0.79 |

**Insights**:

From the above table, comparing the model performance evaluators for the three models, it can be seen that Random Forest model is performing well as compared to the other two models i.e CART and ANN because Random Forest has high precisions for both training and testing data. Hence, Choosing Random Forest Model is the best option in this case as it will exhibit very less variance as compared to a CART or ANN. Also, Random forest provides a very powerful algorithm which results in good predictive accuracy and also reduces instability and between-tree correlation. ANN models have lesser interpretability.

**1.5. To provide business insights and recommendations.**

**1. Business Insights –**

For the business problem of an Insurance firm providing Tour Insurance, we have attempted to make a few Data Models for predictions of probabilities. The models that are attempted are namely, CART or Classification and Regression Trees, Random Forest and Artificial Neural Network. The three models are then evaluated on training and testing datasets and their model performance scores are calculated.

The Accuracy, Precision, F1 Score are computed using Classification Report. The confusion matrix, AUC\_ROC Scores and ROC plot are computed for each model separately and compared. All the three models have performed well but to increase our accuracy in determining the claims made by the customers we can choose the Random Forest Model. Instead of creating a single Decision Tree it can create a multiple decision trees and hence can provide the best claim status from the data.

As seen from the above model performance measures, for all the models i.e. CART, Random Forest and ANN have performed exceptionally well. Hence, we can choose either of the models but choosing Random Forest Model is a great option as even though they exhibit the same accuracy but choosing Random Forest over Cart model is way better as they have much less variance than a single decision tree.

**2. Business Recommendations -**

i) Maximum claims have been raised for the C2B Agency. This probably indicates that a closer look needs to be taken at the client list for the C2B Agency.

ii) The highest claims are for the Silver Plan followed by the Gold plan. The lowest claims are for the Cancellation Plan. This would probably indicate that the Cancellation plan features can be adopted / customized for the Gold and Silver plans too.

iii) The claims for Airlines are roughly 3 times that of the Travel Agency. This would probably mean that a further analysis for Airlines should be carried out to understand why the claim rate is significantly higher.

iv) ASIA accounts for the highest number of sales and also the highest number of claims.

v) The C2B agency operates only in ASIA. The C2B Agency has the highest number of sales. Sales across ASIA are highest for all agencies while EUROPE constitutes for the lowest number of sales. This probably indicates that better offers and increased marketing can be directed for Europe and the Americas.

vi) C2B agency earns the highest commissions followed by CWT. EPX is doing the least business. This probably indicates that better training needed for EPX employees.

vii) The maximum customers for this travel agency are in the age bracket of 30-35years. This could probably mean that there is scope to increase customers in older age brackets by introducing better packages, schemes and plans for the senior citizens.