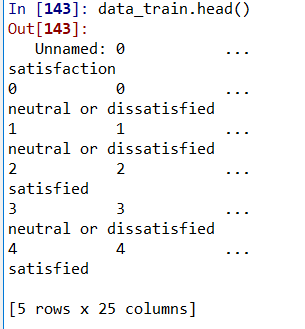
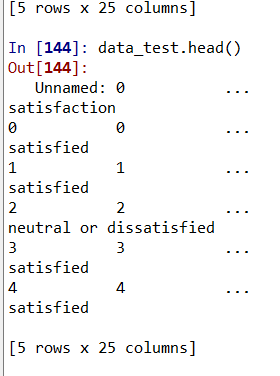
PREDICTING CUSTOMER SATISFACTION RATE

The dataset provided for the assessment contains a lot of information regarding the Customer and their convenience. I have been asked to do data analysis on the dataset and use a machine learning algorithm to identify the most important variables for customer’s satisfaction.

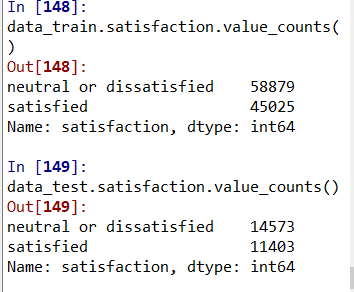
EXPLORATORY ANALYSIS:

AS the dataset was given in two separate files, I have to load them separately first. Names of these two files are train and test. Training file is much bigger than testing file and so does the data within these two files respectively.

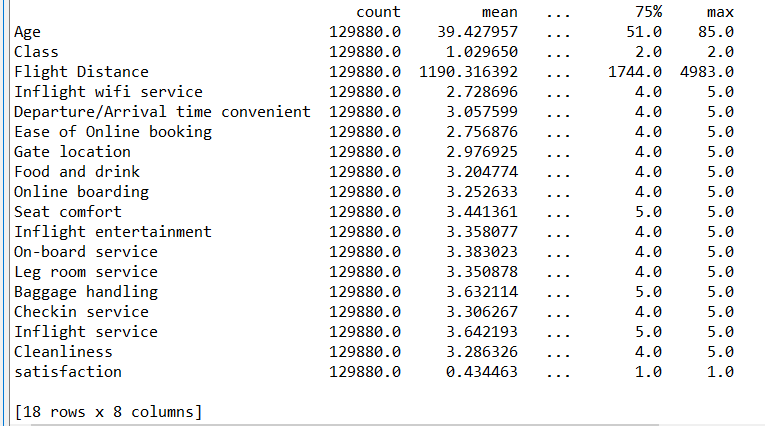
1. After loading these files, I used head() to check the contents of the files.



As you can see both the files have 5 rows and 25 columns. The total columns are 25 but the total rows are 103904

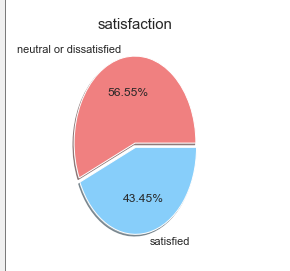
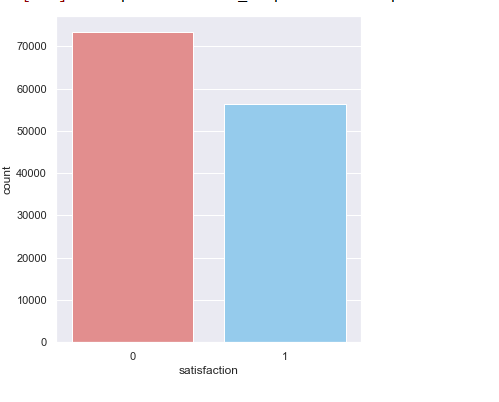
1. 

From the above image, I tried to dig deep and find out more about the satisfaction (of the Customers) and found out that there are total 58879 neutral or dissatisfied customers and 45025 satisfied customers in the Train Dataset.

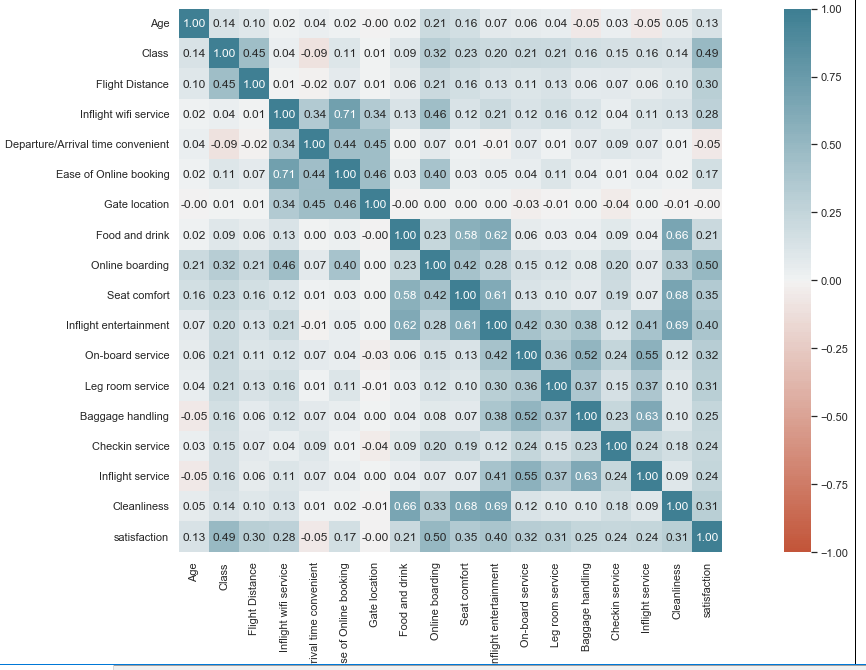
1. 

In this step, I joined both TRAIN and TEST datesets into one and the above image is the description of the one whole dataset and I named it **data** which generally describes it and also shows the mean, median, mode and max of each column separately.

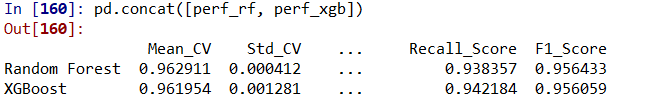
1. After the above step, I tried to find any duplicate values in the dataset but there were none of it.



The above bar graph describes the total number of satisfied and neutral or unsatisfied customers. And the pie-chart explains that there are total 43.45% customers satisfied and 56.55% customers were unsatisfied. This column is also the target column. From the graph, it is shown that there isn’t much of imbalance between satisfied and unsatisfied customers.

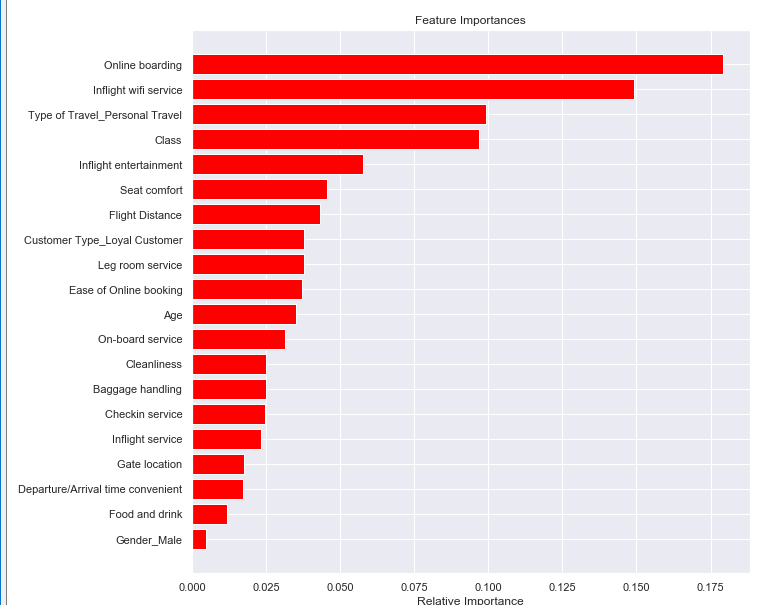
1. 

"Inflight service" and "Baggage handling" are positive correlated with ratio 0.63. "Cleanliness" and "Food and drink" are positive correlated with ratio 0.66.

1. 

The above image describes that I used two models – Random forest and XGboost, both the models performed well but Random forest beat the other model with very slight difference if I compare them using their F1- Score.

F1- Score : F1 score is a part of confusion matrix which depicts the balance between precision and recall.



The above image shows the most important variables which were online boarding, Inflight- wifi services and Type of travel. In the end, it was concluded that Gender plays zero role in customer’s satisfaction. Generally older customers around age 30 to 60 are likely to be more satisfied.