Data Science Additional Project

support@intellipaat.com

Problem Statement -

The objective is to analyze the flight booking dataset obtained from a platform which is used to book flight tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers. Apply EDA, statistical methods and Machine learning algorithms in order to get meaningful information from it.

Dataset -

Flight booking dataset contains 300152 records and 11 features for the flight travel between India's top 6 metro cities.

Dataset Description-

Attributes	**Description**
Airline	Name of the airline company
Flight	Plane's flight code
Source City	City from which the flight takes off
Departure Time	Time of Departure
Stops	Number of stops between the source and destination cities
Arrival Time	Time of Arrival
Destination City	City where the flight will land
Class	Contains information on seat class
Duration	Overall amount of time taken to travel between cities in hours.
Days left	Subtracting the trip date by the booking date.
Price	Ticket price

1) Plot the best suitable graph if the price varies with different airlines.

2) Is the price affected when tickets are bought just 1 or 2 days before departure? Show the stats in graphical form.

3) Which airline has been used by the most number of passengers ?

4) How does the ticket price vary between Economy and Business class for all airlines? Show the insights with the best suitable graph.

5) How many passengers are travelling in business class on a direct flight ?

6) Which city has the lowest take off of flights ?

7) What is the average price of all airlines from Delhi to Mumbai ?

8) Convert all the attributes into integer data type which are of object data type except Flight column.

9) Which feature is highly dependent on the target variable ?

Perform the following tasks for answering the remaining questions -

- Use airline, class, duration and days left as a feature.
- Split the dataset into a train and test of proportions 80:20 and set the random state to 0.
- Build a Multiple Linear regression model on the data.

Answer the following questions with the help of the above-created model.

10) Calculate the MSE, MAE, R-squared and RMSE.

11) Plot the graph of Actual and Predicted Values

12) Implement Decision Tree and Random forest algorithm. Which one is best algorithm out of the three algorithms used on this dataset ?