Flight Price Prediction

Overview

The objective of this article is to predict flight prices given the various parameters. We need to train a model over a given dataset. The algorithm that we will be using would be a regression algorithm as the target variable that is, “Price” is continuous data.

Introduction

Airline companies use complex algorithms for calculating the price of the ticket for a particular time to maximize the revenue.

We can ourselves observe that the number of people using airlines to travel to their destinations is drastically increasing. It is difficult for an airline to maintain the same price as they have various factors to consider, that’s why we use a machine learning algorithm to help the airline to predict the prices.

1. Problem Definition

Flight ticket prices are something hard to guess, today we might see one price but tomorrow we might see another. We often hear travelers saying that the price of the tickets is very unpredictable.

Here we are provided with prices of flight tickets for various airlines between the months March 2019 to June 2019 between various cities.

The problem statement explains that the target variable is continuous and it’s a “Regression Type Problem” since we need to predict the price of flight tickets. In this project we will be using various regression models to help the customers to make their purchasing decision accordingly.

Attribute Information:

**Airline:** The name of the airline

**Date\_Of\_Journey:** The date of the journey

**Source:** The source from where the service begins

**Destination:** The destination where the service ends

**Route:** The route taken by the flight to reach the destination

**Dep\_Time:** The time the journey starts from the source

**Arrival\_Time:** The time of arrival at the destination

**Duration:** Total duration of the flight

**Total\_Stops:** Total stops between the source and destination

**Additional\_Info:** Additional information about the flight

**Price:** The price of the ticket.

The flight price prediction dataset has a total of 11 attributes with its description where “**Price**” is the target variable, whereas the other 10 attributes are the independent variables.

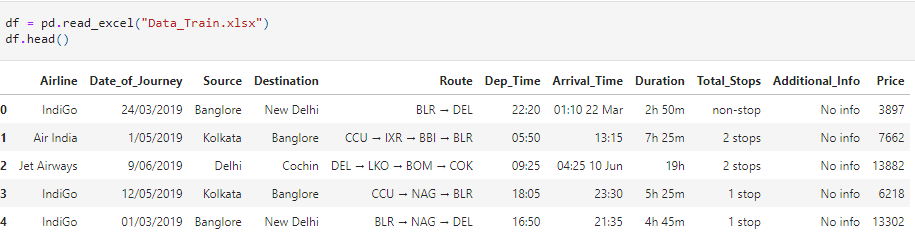
1. Data Analysis

Data analysis is the process where the data is examined and conclusions or insights are drawn from it.

There are 2 dataset that have been provided; one is the training dataset and the other is the testing dataset.

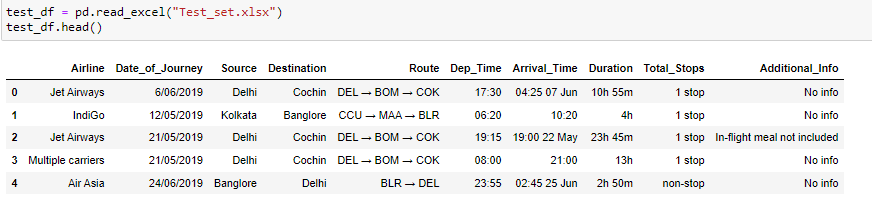
Train dataset: The training dataset is used to train the model. There are 10 independent variables and 1 dependent variable.

The training dataset has 10,638 records.



Test dataset: The test dataset has only the independent variables.

The testing dataset has 2671 records.



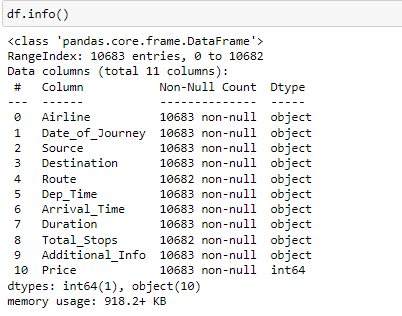
Understanding the Dataset

There are 4 attributes that need to be modified.

1. Date\_of\_Journey: This attribute has the date when the journey started, but our models do not understand this kind of data, so we need to extract the day, month and year from this attribute and then drop the column.
2. Dep\_Time: This attribute has a combination in hours: minutes, we need to extract them individually and then drop the column.
3. Arrival\_Time: This attribute has both the date and time of arrival, we need to separate them into individual columns and then drop the column.
4. Duration: This attribute contains the duration of the flight, we need to convert the entire duration into minutes.

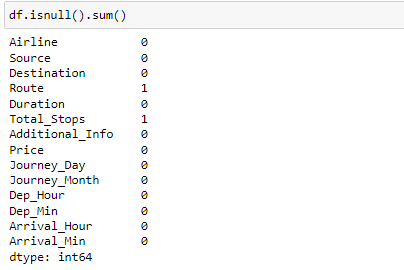
Exploratory Data Analysis (EDA)

EDA is the first step in the process of data analysis. Let us see the summary of the dataset. The info() method in the pandas library gives us the summary of the dataset.



Checking Missing Values

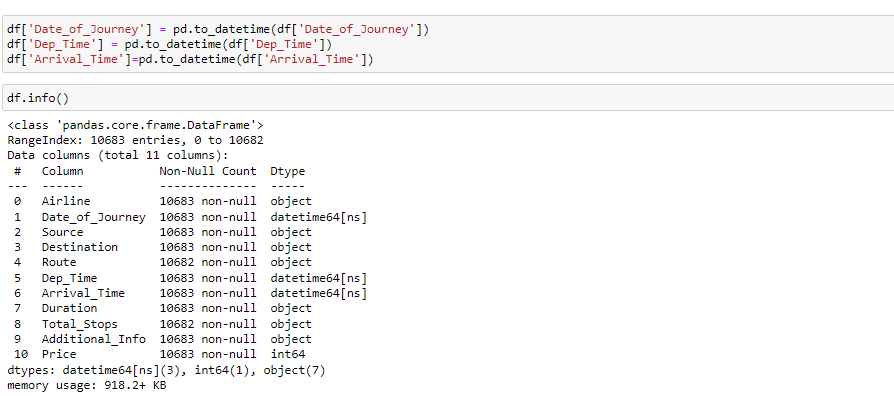
We can check if there are any missing values in the dataset using the isnull().sum() method, this will tell us the total count of missing values in each attribute in our dataset.



From the above image we can see that the attributes Route and Total\_stops have 1 missing value each.

Feature Engineering

Feature engineering is the process of selecting and transforming the relevant variables from raw data when creating a model. The columns Date\_of\_Journey, Arrival\_Time, and Dep\_Time has to be in the datetime format.



We can see that we have converted the object type of data in the above mentioned columns to datetime type. We can now extract the data from these columns.

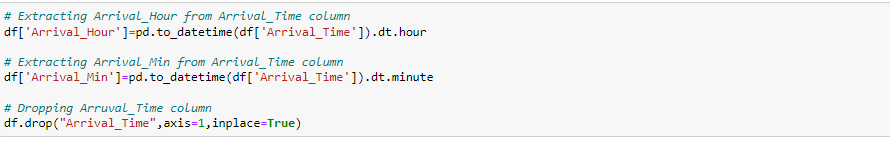
**Date\_of\_Journey:** This attribute has the date when the journey started, but our models do not understand this kind of data, so we need to extract the day, month and year from this attribute and then drop the column.



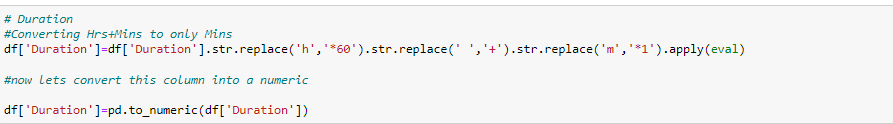
**Dep\_Time:** This attribute has a combination in hours: minutes, we need to extract them individually and then drop the column.



**Arrival\_Time:** This attribute has both the date and time of arrival, we need to separate them into individual columns and then drop the column.



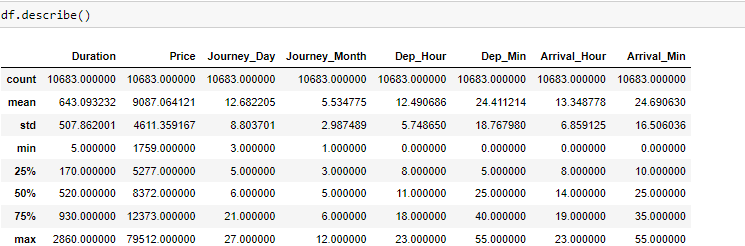
**Duration:** This attribute contains the duration of the flight, we need to convert the entire duration into minutes.



There are few values in the dataset which have repeated names but in different formats, let us change them to one format.

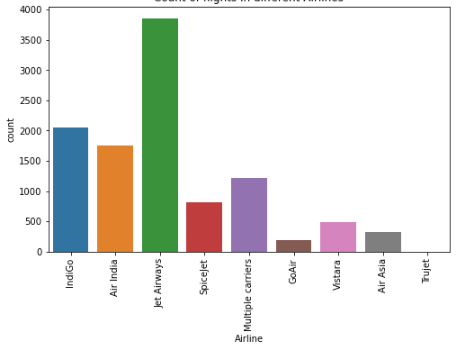


Description of the dataset

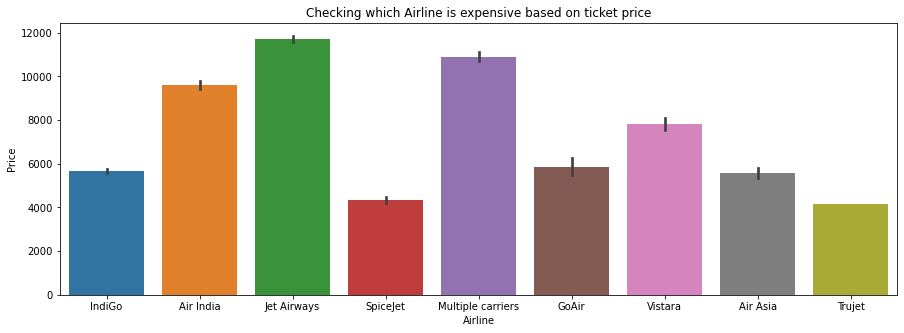


Data Visualization

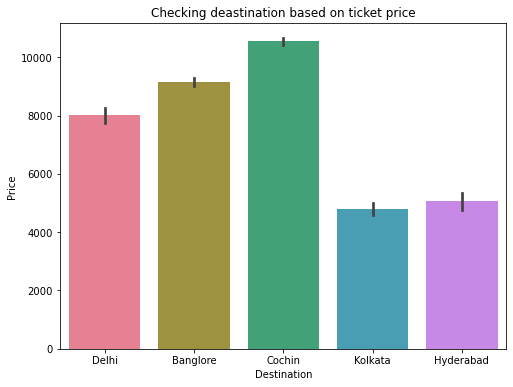
Let us visualize the categorical columns.



From the above figure we can see that the number of flight Jet airways operates is higher when compared to the other flight companies.

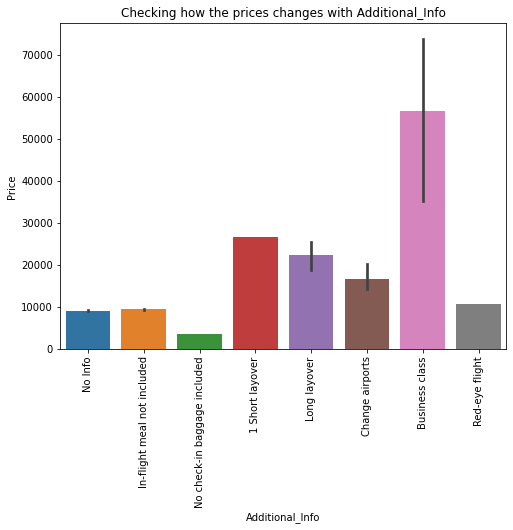


From the above image we can observe that the “Jet Airways” has the highest price range.

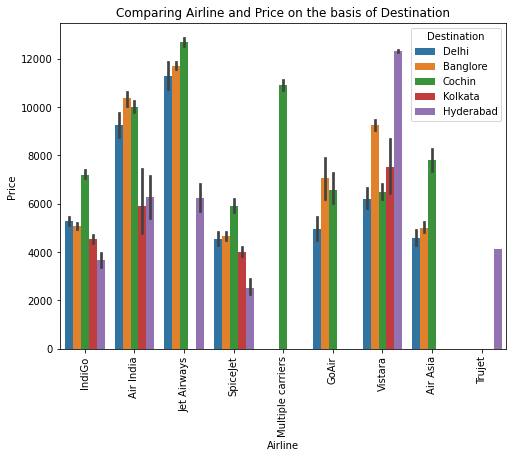


We can observe the following from the above graph:

1. When the destination city is Cochin, the price range of the tickets is the highest
2. The price range of the tickets is the least when the destination city is Kolkata

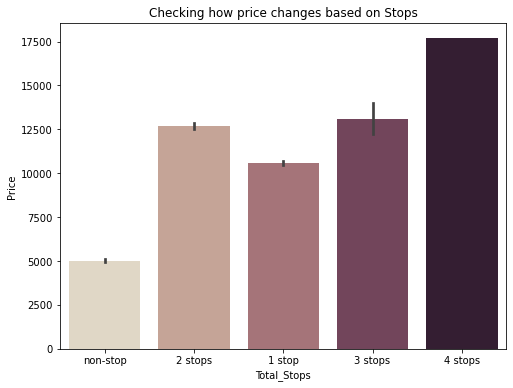


We can see that Business class tickets cost the most as they are the most luxurious.



We can observe the following from the above graph:

1. Jet Airways flights with the destination as Delhi have the highest ticket price.
2. Indigo flights with the destination as Cochin has the highest price.
3. Air India flights with the destination as Bangalore have the highest ticket price.
4. Spice Jet ticket price is the highest when the destination is Cochin
5. Go Air ticket prices are almost the same when the destination is Cochin/Bangalore
6. Vistara ticket prices is the highest when the destination is Hyderabad
7. Air Asia ticket price is the highest when the destination is Cochin

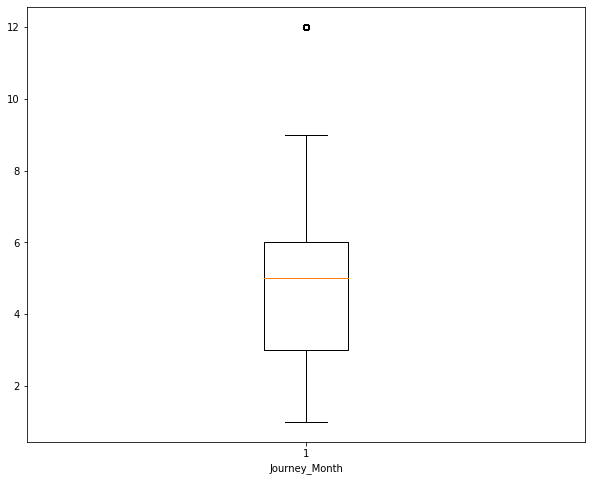
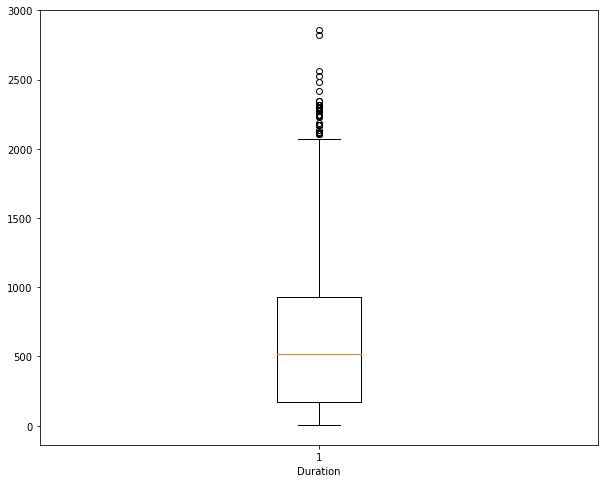


We observe that if the number of stops is high the price range of the ticket is higher as well.

Checking Outliers

We can check if there are outliers by plotting a boxplot. We check outliers only in numerical columns.

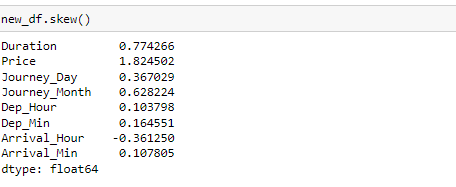


We observe after running the above code, that the features Duration and Journey\_Month have outliers, so we remove the outliers using Z-score method.

Checking Skewness

Skewness tells us the measure of deviation from the normal distribution is known as skewness.

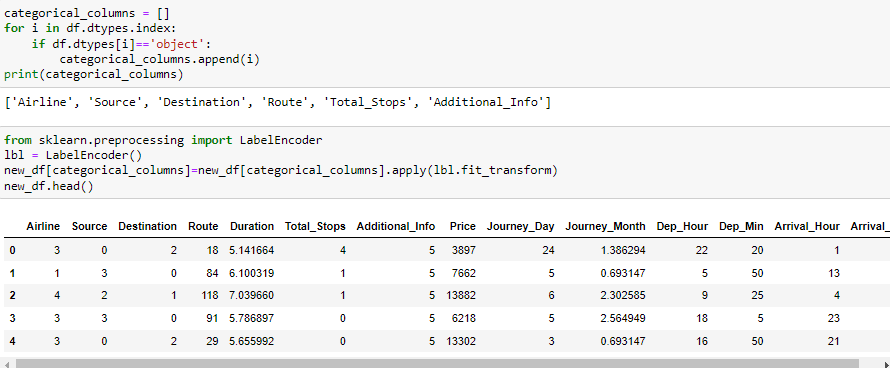


If the skew is above or below +/- 0.5, we need to reduce the skewness by applying some methods. In my project I used “log transformation” method to remove the skewness.

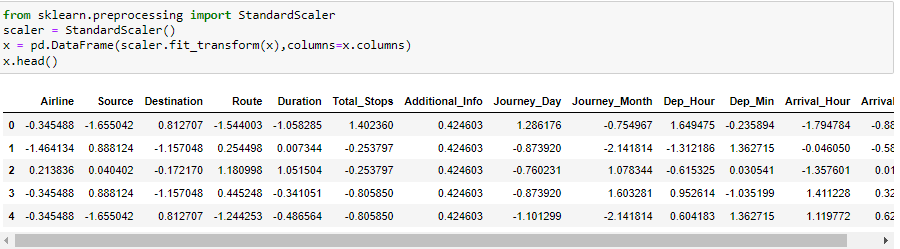
1. EDA Concluding Remarks
2. The number of flights in the month of January is the highest when compared to all the other months
3. Most of the flights start for the capital city, Delhi while Cochin seems to be the destination of maximum number of flights
4. The price of the ticket is the highest when the flight has 4 stops
5. Business class tickets are the most expensive
6. Data Pre-processing

The process of transforming the raw data for using them in the algorithms.

1. Label Encoding: This step is done to convert all the categorical data into numerical data.

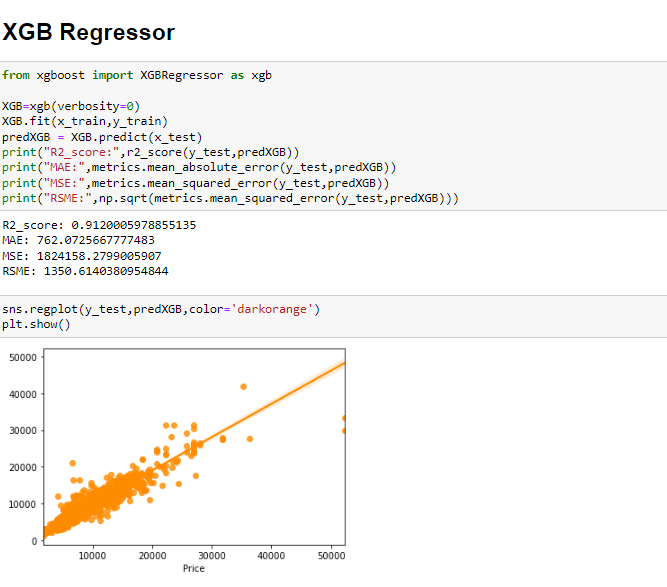


1. Standard Scaler: It is the process of scaling the data such that the mean is 0 and standard deviation is 1.

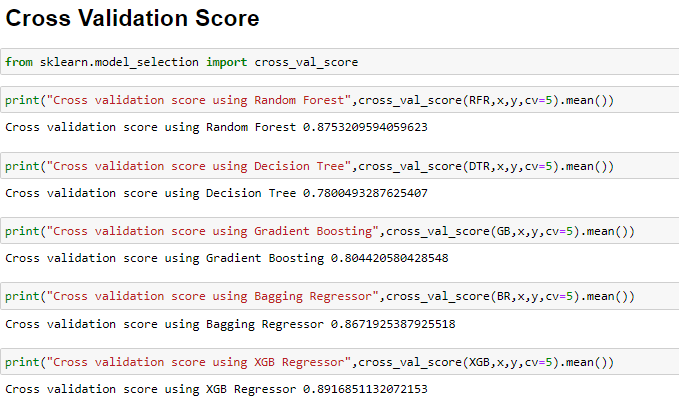


1. Model Building

This is the main step of machine learning, fitting the model and predicting the results. Here, we use multiple different models and then select the best model. I found the model to be XG Boost Regressor, which gave me an **R2 score** of **91.2%.**

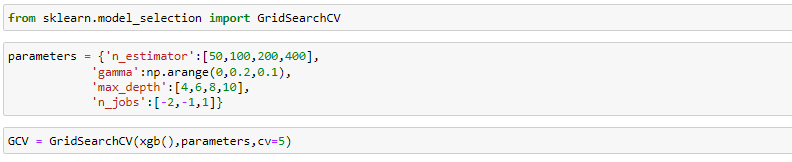


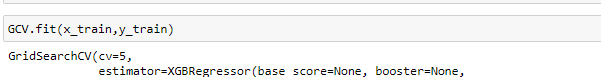
Let us now check the cross validation score and check for overfitting.

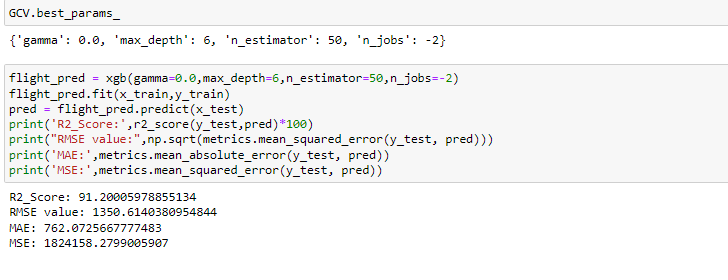


We can see that we got the highest accuracy score using **XG Boost Regressor**.

Hyper parameter Tuning







We have now hyper tuned our model, let us now save our model using Pickle.

1. Conclusion





We have successfully trained the regression model **“XG Boost Regressor”** to predict the price of flight tickets with an r2\_score of **91.2%.**