FLIGHT PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES

**DILIP KUMAR**

**DATA TRAINED:B.NO-1829**

**DEPARTMENT OF B.E(ENTC)**

**ACEM,UNIVERSITY OF PUNE,INDIA**

**EMAIL-ID:DK4U90@GMAIL.COM**

**Abstract:**

Someone who purchase flight tickets frequently would be able to predict the right time to procure a ticket to obtain the best deal. Many airlines change ticket prices for their revenue management. The airline may increase the prices when the demand is to be expected to increase the capacity. To estimate the minimum airfare, data for a specific air route has been collected including the features like departure time, arrival time and flight distance, purchasing time, fuel price, etc. Features are extracted from the collected data to apply Machine Learning (ML) models. Each carrier has its own proprietary rules and algorithms to set the price accordingly. Recent advance Machine Learning (ML) makes it possible to infer such rules and model the price variation. This paper gives the machine learning regression methods to predict the prices at the given time.

Applied Lazy Prediction for regressor Module Selection .

**Contents:**

Business Understanding …………………………………………………………….1

Data Understanding …………………………………………………………………2

Data Preparation …………………………………………………………………….3

Importing modules…………………………………………………………3.1

Visualization………………………………………………………………..3.2

Handle null values …………………………………………………………3.3

Data Cleaning………………………………………………………………3.4

Data Exploration and Transformation……………………………………...3.5

Split dataset…………………………………………………………………3.6

Feature Importance………………………………………………………….3.7

Normalization …………………………………………………………….3.8

Modeling ………………………………………………………………………….….4

Evaluation ………………………………………………………………………...….5

Save model ……………………………………………………………………………6

End Notes ……………………………………………………………………………7

1. **Business Understanding:**

Since the deregulation of the airline industry, airfare pricing strategy has developed into a complex structure of sophisticated rules and mathematical models that drive the pricing strategies of airfare. Although still largely held in secret, studies have found that these rules are widely known to be affected by a variety of factors. Traditional variables such as distance, although still playing a significant role, are no longer the sole factor that dictate the pricing strategy. Elements related to economic, marketing and societal trends have played increasing roles in dictating the airfare prices. Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination. Being able to predict the airfare trend at the specific market segment level is crucial for airlines to adjust strategy and resources for a specific route. However, existing studies on market segment price prediction use heuristic-based conventional statistical models, such as linear regression, and are based on the assumption that there exists a linear relationship between the dependent and independent variables, which in many cases, may not be true. Recent advances Machine Learning (ML) make it possible to infer rules and model variations on airfare price based on a large number of features, often uncovering hidden relationships amongst the features automatically. To the best of our knowledge, all existing work leveraging machine learning approaches for airfare price prediction are based on:

1. proprietary datasets that are not publicly available and
2. transaction records data crawled from online travel booking sites like Kayak.com .

The problem of the former lies in the difficulty of gaining access to the data, making reproducing the results and extending the work nearly impossible. The issue with the later is that the transaction records from each online booking site are a small fraction of the total ticket sales from the entire market, making the acquired data likely to be skewed, and thus, not representing the true nature of the entire market. In this paper, we address the problem of market segment level airfare price prediction by using publicly available datasets and a novel machine learning framework to predict market segment level airfare price. More specifically, our proposed framework extracts information from two specific public datasets, that are collected and maintained by the Office of Airline Information within the United States Bureau of Transportation Statistics (BTS).

1. **Data Understanding:**

Air ticket price prediction is a challenging task since the factors involved in pricing dynamically change over time and make the price fluctuate. In the last decade, researchers have incorporated machine learning algorithms and data mining strategies to better model observed prices. Among them, regression models, such as Linear Regression (LR), Support Vector Machines (SVMs), Random Forests (RF), are frequently used in predicting accurate airfare price .

Before starting data preparation let’s have a glimpse of data first.

As we saw in Data Analysis there are 11 variables in the given data. Below is the description of each variable.

**Airline:** Name of the airline used for traveling

**Date\_of\_Journey:** Date at which a person traveled

**Source:** Starting location of flight

**Destination:** Ending location of flight

**Route:** This contains information on starting and ending location of the journey in the standard format used by airlines.

**Dep\_Time:** Departure time of flight from starting location

**Arrival\_Time:** Arrival time of flight at destination

**Duration:** Duration of flight in hours/minutes

**Total\_Stops:** Number of total stops flight took before landing at the destination.

**Additional\_Info:** Shown any additional information about a flight

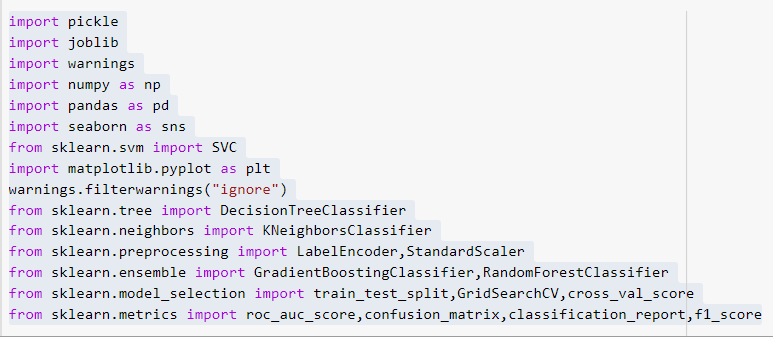
**Price:** Price of the flight

Few observations about some of the variables:

1. ‘Price‘ will be our dependent variable and all remaining variables can be used as independent variables.

2. ‘Total\_Stops‘ can be used to determine if the flight was direct or connecting.

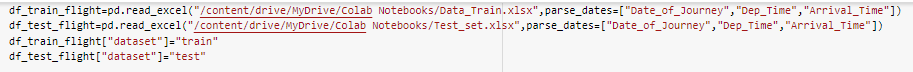
1. **Data Preparation:**
   1. **Importing modules:**

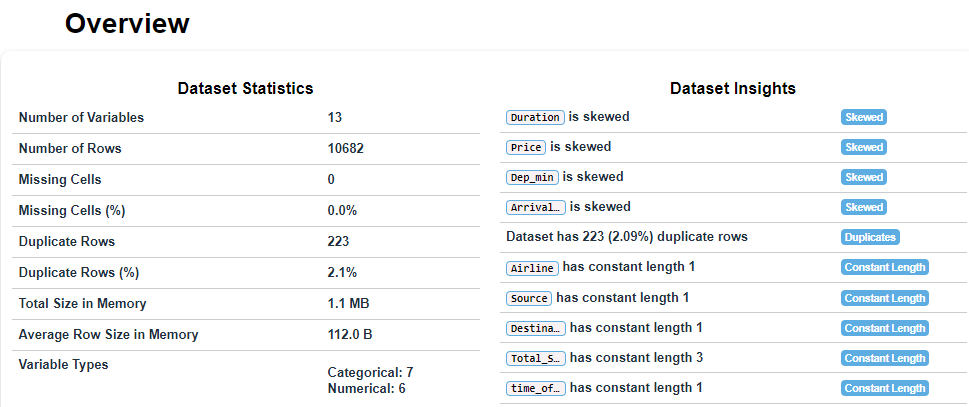
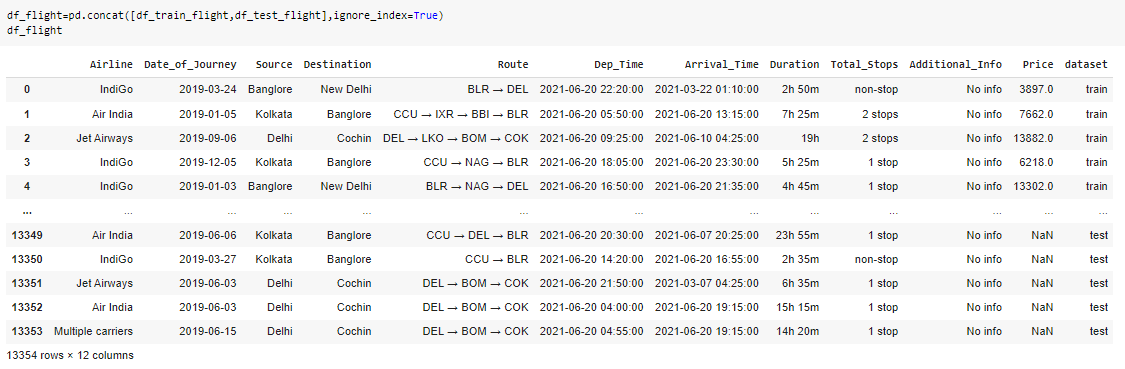


Let’s we take brief about these libraries, pandas are used for data analysis NumPy is for n-dimensional array seaborn and matplotlib both have similar functionalities which are used for visualization.

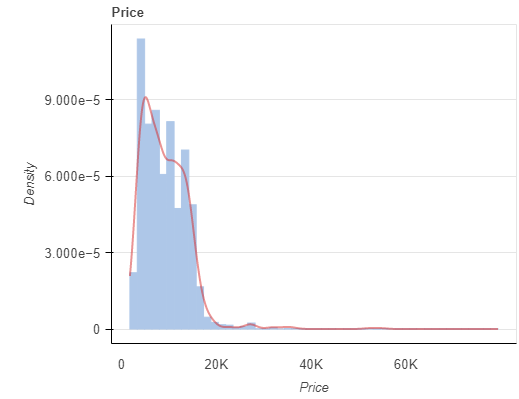
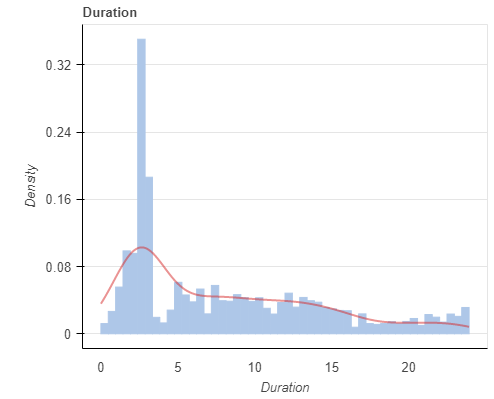
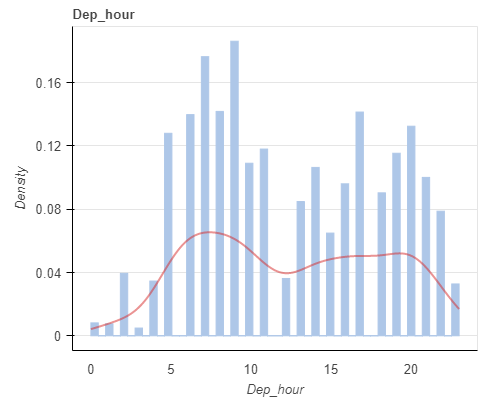
**3.2 Visualization:**

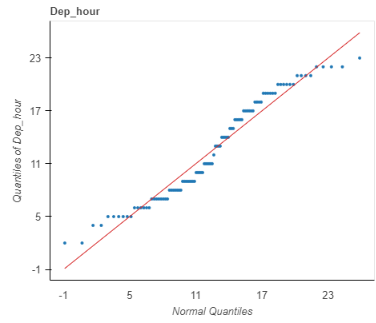
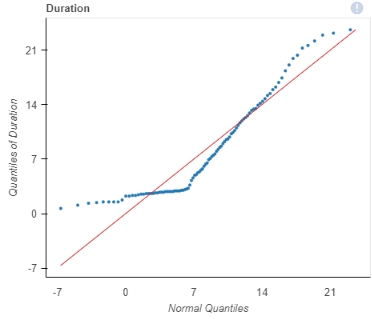
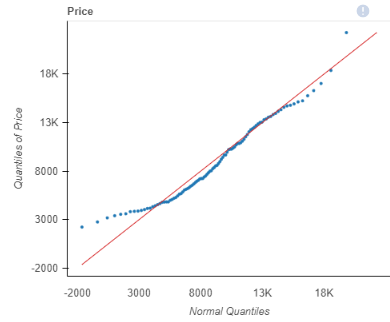
We know that the “image speaks everything” here the visualization came into the work, we use visualization for explaining the data. In other words, we can say that it is a graphic representation of data that is used to find useful information.



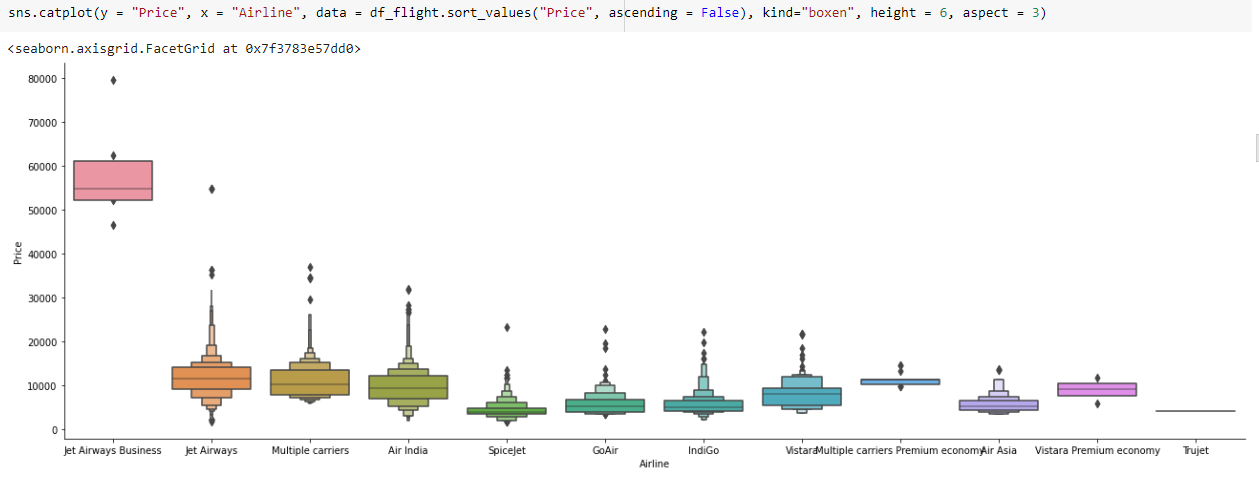
Output

Next, for independent numerical variables, the first step to further analyze the relationship with our dependent variable was to create density plots visualizing the spread of the data

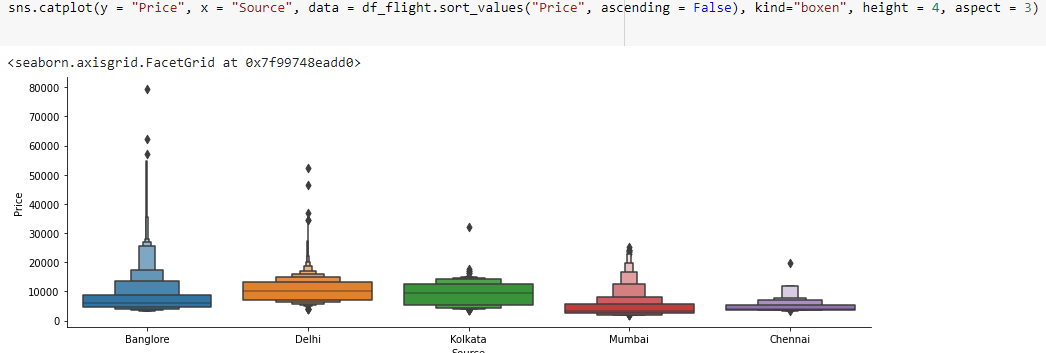
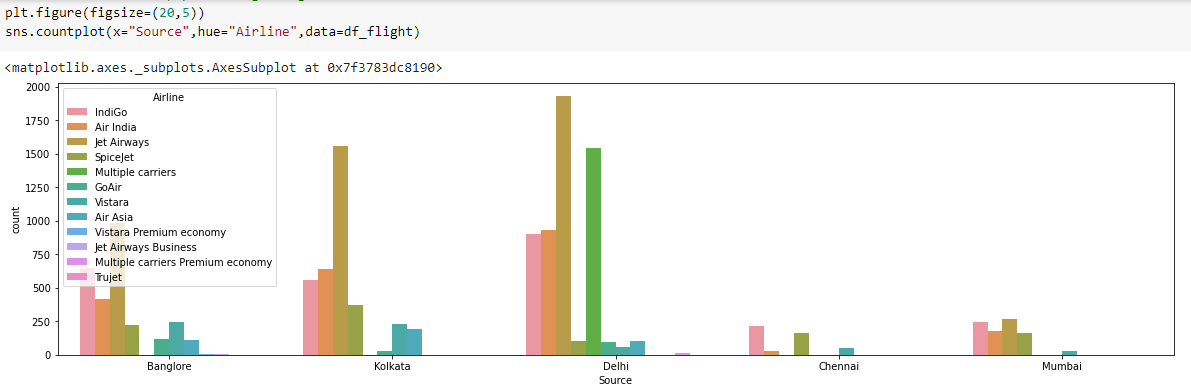


Let’s see how the Airline variable is related to the Price variable.



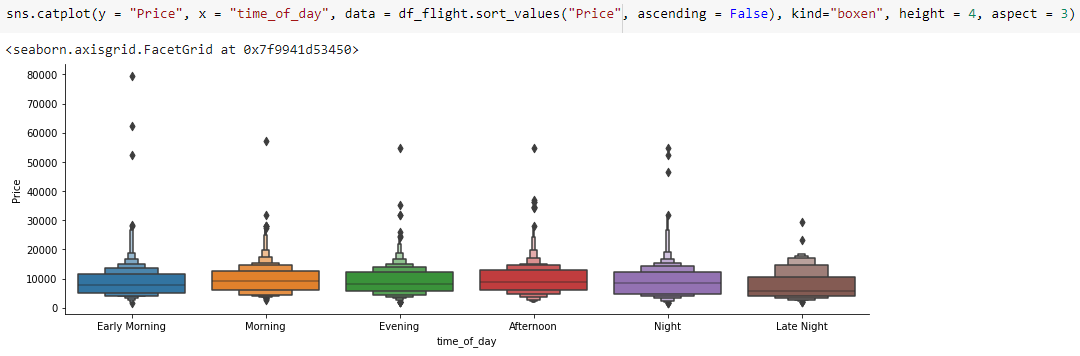
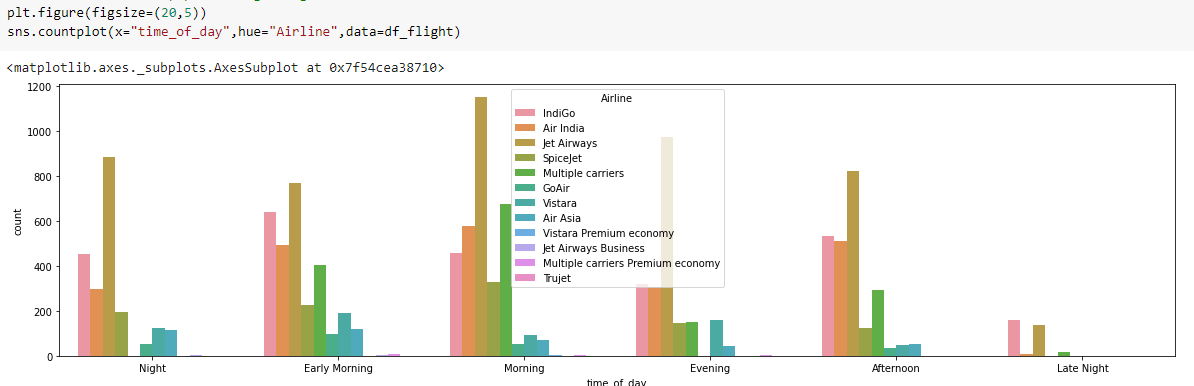
As we can see the name of the airline matters. ‘JetAirways Business’ has the highest price range. Other airlines price also varies.

Let’s see how the Airline variable is related to the source variable.



As we can see the source matters. ‘Banglorehas the highest price range comperision to Others source.

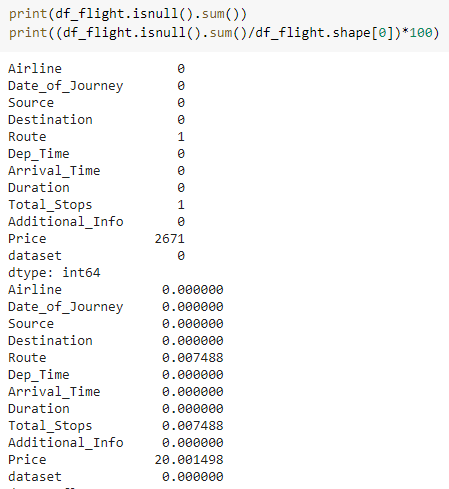
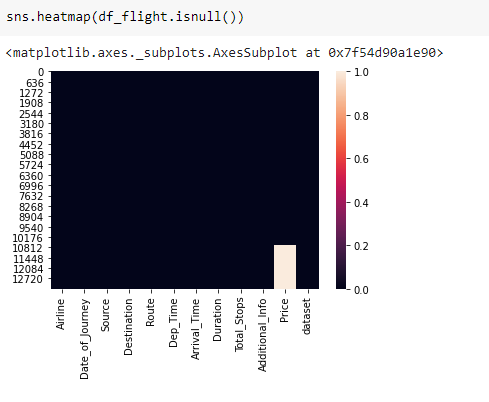
Let’s see how the Airline variable is related to the day of time variable.



As we can see the time of the day matters. ‘Early Morning’ time has the highest price range Other time.

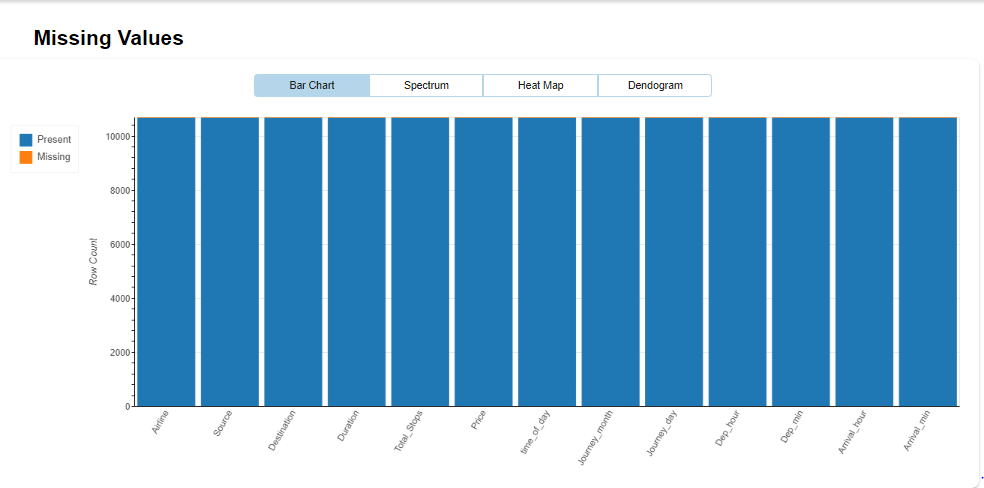
**3.3 Handling Missing Values:**

As we found out the ‘Route’ and ‘Total\_Stops’ variables have very low missing values in data. Let’s now see the percentage of missing values in data

As we can observe ‘Route’ and ‘Total\_Stops’ both have 0.0074% of missing values. In this case, it is better to drop missing values.

After fill and remove null values final data set has no null Value is present.



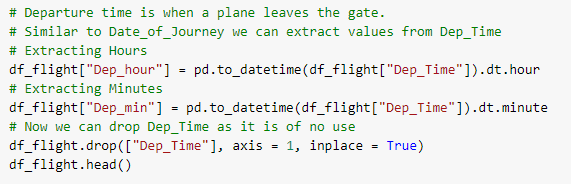
Now we don’t have any missing values.

**3.4 Data Cleaning:**

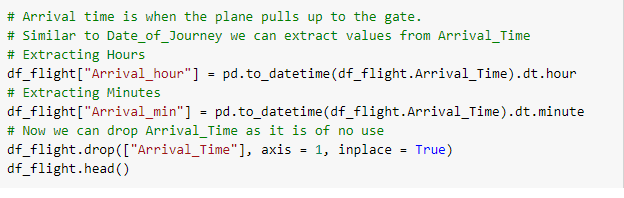
**Handling Date and Time Variables**

We have ‘Date\_of\_Journey’, a ‘date type variable and ‘Dep\_Time’, ‘Arrival\_Time’ that captures time information.

We can extract ‘Journey\_day’ and ‘Journey\_Month’ from the ‘Date\_of\_Journey’ variable. ‘Journey day’ shows the day of the month on which the journey was started.

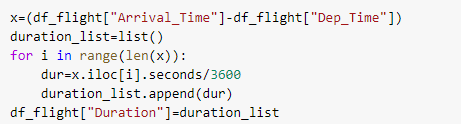


Similarly, we can extract ‘Departure\_Hour’ and ‘Departure\_Minute’ as well as ‘Arrival\_Hour and ‘Arrival\_Minute’ from ‘Dep\_Time’ and ‘Arrival\_Time’ variables respectively.



We also have duration information on the ‘Duration’ variable. This variable contains both duration hours and minutes information combined.

We can also calculate Duration to subtract from Arrival time to Departure Time.



**Handling Categorical Data**

Handling Categorical Data One can find many ways to handle categorical data. Some of them categorical data are,

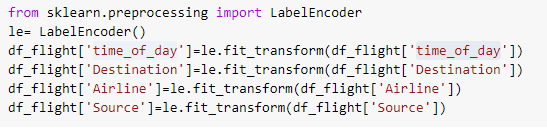
Nominal data --> data are not in any order --> OneHotEncoder is used in this case

Ordinal data --> data are in order --> LabelEncoder is used in this case

Airline, Source, Destination, Route, Total\_Stops, Additional\_info are the categorical variables we have in our data. Let’s handle each one by one.

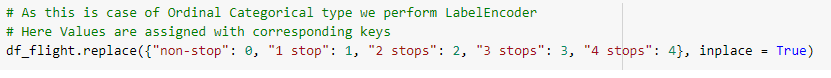
Since the Airline variable is Nominal Categorical Data (There is no order of any kind in airline names) we will use LableEncoder to handle this variable.

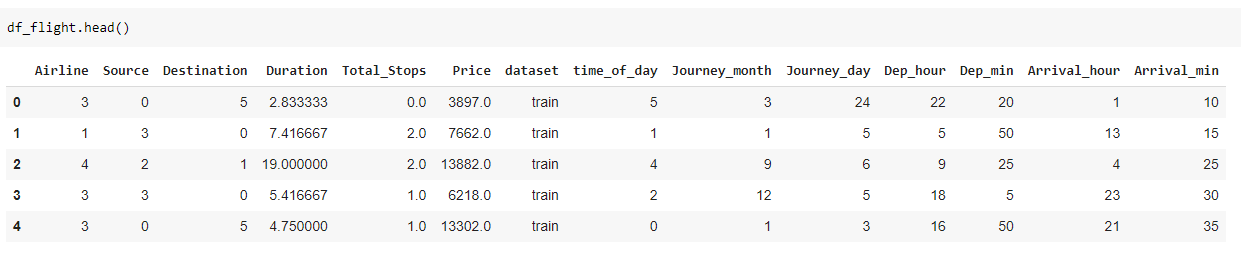
**LableEncoder** as shown in the above code.



Total\_Stops Variable

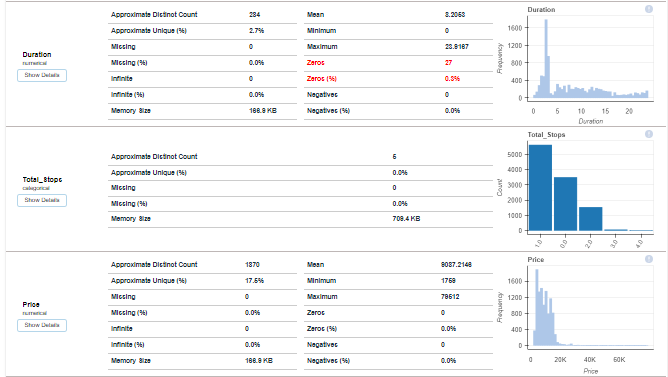
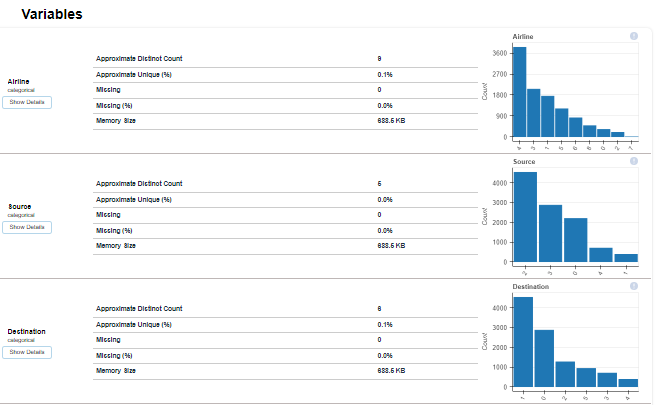
Here, non-stop means 0 stops which means direct flight. Similarly meaning other values is obvious. We can see it is an Ordinal Categorical Data so we will use LabelEncoder here to handle this variable.

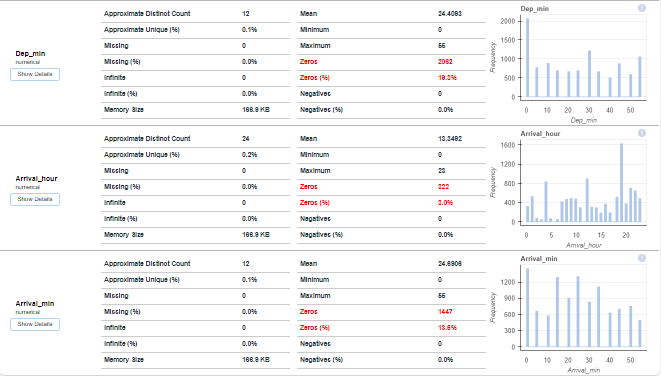
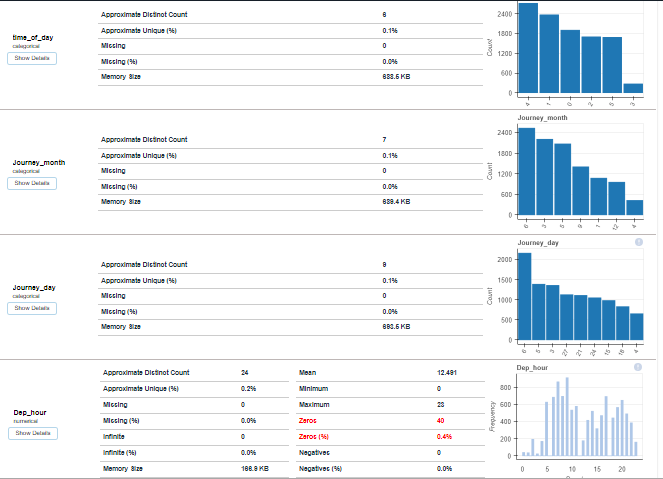
Now we will create the final dataframe by concatenating all the One-hot and Label-encoded features to the original dataframe. We will also remove original variables using which we have prepared new encoded variables.

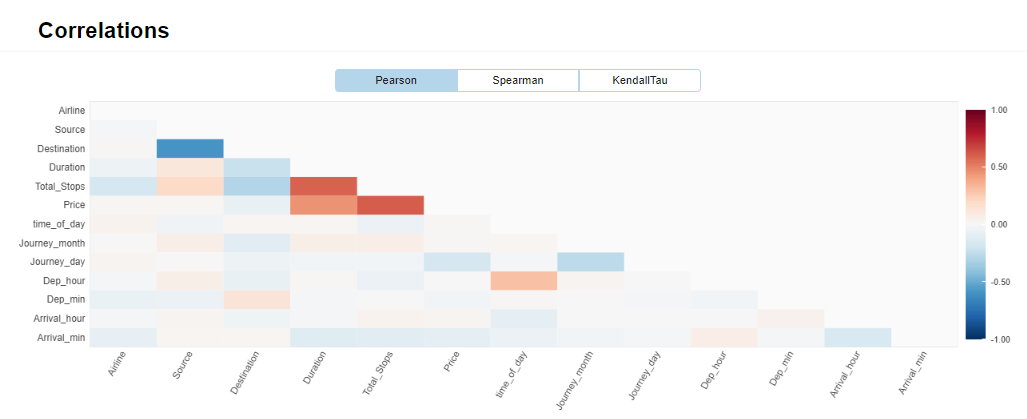


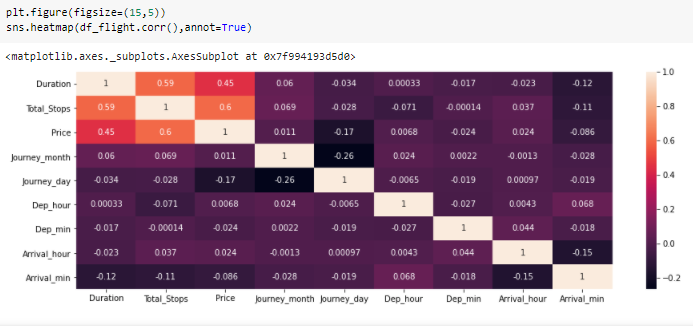
### 3.5 Data Exploration and Transformation:

Data Exploration and Transformation To see which variables are likely to affect the flight price the most, I ran a correlation analysis of our independent variables against our dependent variable, quality. This analysis ended up with a list of variables of interest that had the highest correlations with quality.





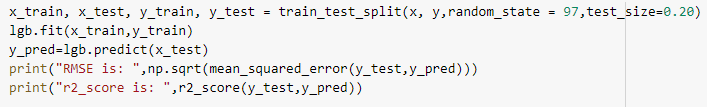
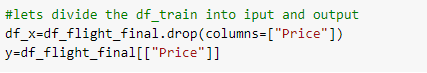




A heat map is a two dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The seaborn python package allows the creation of annotated heatmaps which can be tweaked using matplotlib tools as per the creators requirement.

**3.6 Split dataset:**

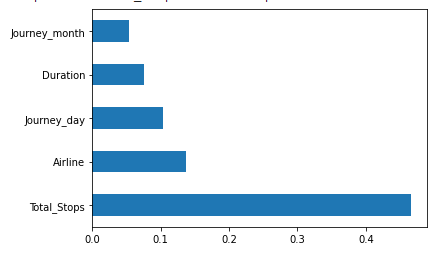
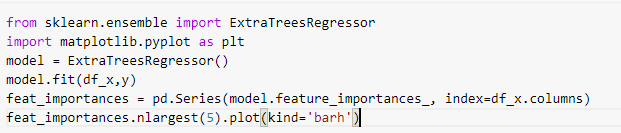
Now we perform a split operation on our dataset.



**3.7 Feature Importance:**

Another great quality of random forest is that they make it very easy to measure the relative importance of each feature. Sklearn measure a features importance by looking at how much the treee nodes, that use that feature, reduce impurity on average (across all trees in the forest). It computes this score automaticall for each feature after training and scales the results so that the sum of all importances is equal to 1. We will acces this below:

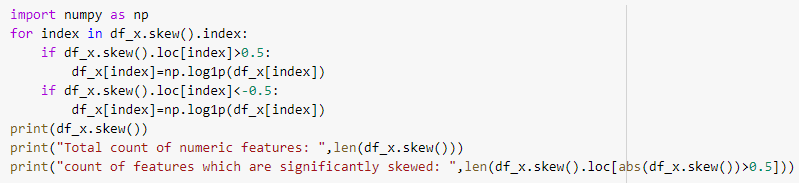
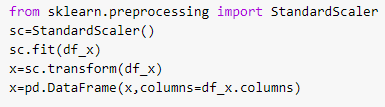
Top Five Feature importance Graphysically representations



**3.8 Normalization:**

My first step was to clean and prepare the data for analysis. I went through different steps of data cleaning. First, I checked the data types focusing on numerical and categorical to simplify the correlation's computation and visualization. Second, I tried to identify any missing values existing in our data set. Last, I researched each column/feature's statistical summary to detect any problem like outliers and abnormal distributions.

We do normalization on numerical data because our data is unbalanced it means the difference between the variable values is high so we convert them into 1 and 0.

1. **Applying Model**

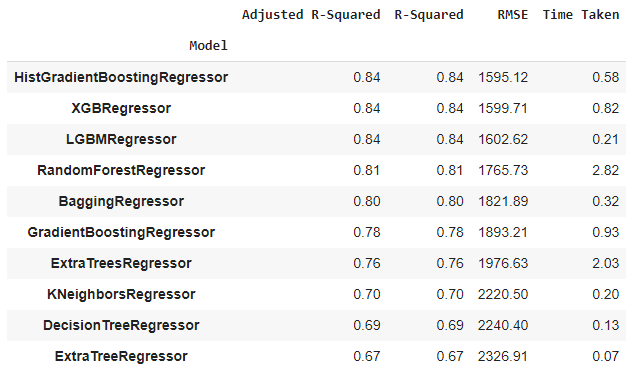
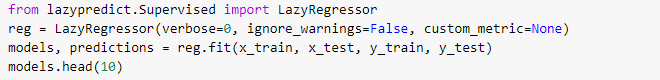
### **Applying Lazy Prediction**

One of the problems of the model-building exercise is ‘How to decide which machine learning algorithm to apply ?’

This is where Lazy Prediction comes into the picture. Lazy Prediction is a machine learning library available in python that can quickly provide us with performances of multiple standard classifications or regression models on multiple performance matrices.

Let’s see how it works…

Since we are working on a Regression task we will use Regressor models.

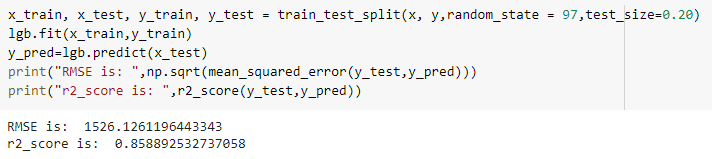


As we can see LazyPredict gives us results of multiple models on multiple performance matrices. In the above figure, we have shown the top ten models.

Here ‘XGBRegressor’ and ‘ExtraTreesRegressor’ outperform other models significantly. It does take a high amount of training time with respect to other models. At this step we can choose priority either we want ‘time’ or ‘performance’.

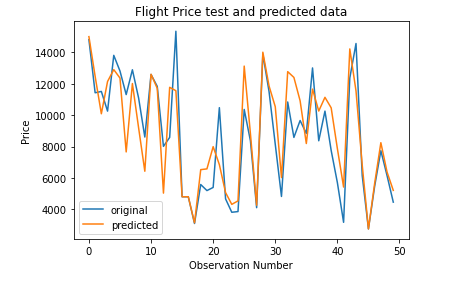
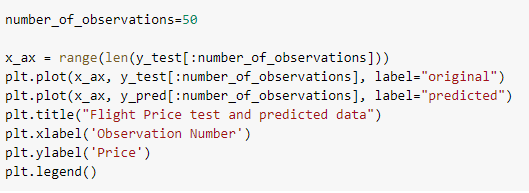
We have decided to choose ‘performance’ over training time. So we will train ‘ExtraTreesRegressor’ and visualize the final results.

**ExtraTreesRegressor:**

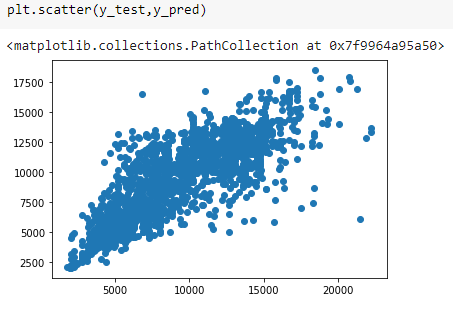
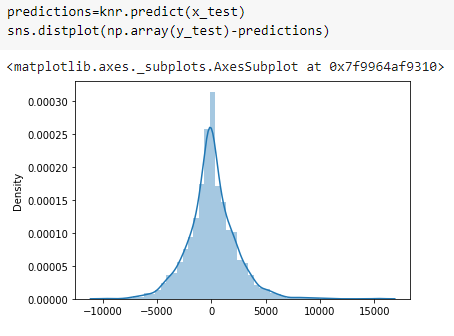


1. **Evaluation:**

Now, we are at the end of our article, we can differentiate the predicted values and actual value.



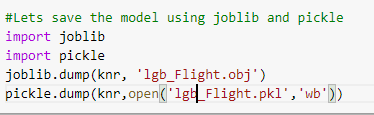
As we can observe in the above figure, model predictions and original prices are overlapping. This visual result confirms the high model score which we saw earlier.

Y\_test and Y\_pred value Linearly distrubuted. Y\_test and Y\_pred value Normally Distrubuted.

### **Saving Model**

At last, we save our machine learning model:



## **End Notes:**

This is one of the interesting articles that I have written because it was on today’s current top technology machine learning, but I was used basic language to explain this article so, you can’t get difficulty on understanding.

A proper implementation of this project can result in saving money of inexperienced people by providing them the information related to trends that flight prices follow and also give them a predicted value of the price which they use to decide whether to book ticket now or later. In conclusion this type of service can be implemented with good accuracy of prediction. As the predicted value is not fully accurate there is huge scope for improvement of these kind of service.