**BLOG SUBMISSION**

**Flight Ticket Price Prediction**



**Submitted by:**

**Gyan Prakash Tripathi**

**ACKNOWLEDGMENT**

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on –

1. Time of purchase patterns (making sure last-minute purchases are expensive)

2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases)

In this blog, I have done data collection of flight ticket price through web scraping from online website **Yatra.com.** Then I have a**nalysed the flight ticket fare prediction using Machine Learning dataset** using essential exploratory data analysis techniques and also, I will be performing some data visualizations to better understand our data.

**In the dataset, I have scraped many columns like Departure time, Arrival time, Duration, source, destination meal info and so on. There are more than 1900 rows in the dataset which gives flight price details of different source and destination cities.**

By doing data preprocessing, data analysis, feature selection, and many other techniques we built our cool and fancy machine learning model. And at the end, we applied many ml algorithms to get the very good accuracy of our model.

**Many thanks to Fliprobo for providing me this project to understand about the Real Time Field work present in Data Science Industry.**

I am very thankful to my friends and family who helped me through this study. So without any further due.

**ABSTRACT**

Now-a-days flight prices are quite unpredictable. The ticket prices change frequently. The price of an airline ticket is affected by a number of factors, such as ﬂight distance, purchasing time, fuel price, etc. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible. Using technology, it is actually possible to reduce the uncertainty of flight prices. Each carrier has its own proprietary rules and algorithms to set the price accordingly. Recent advance in Artiﬁcial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation.

So here we will be predicting the flight prices using efficient machine learning techniques.

**TAKEAWAYS FROM THE BLOG**

In this article, we do prediction using machine learning which leads to the below takeaways:

1. **Web Scraping:** Scraping data from websites like Yatra.com
2. **EDA:** Learn the complete process of EDA
3. **Data analysis:** Learn to withdraw some insights from the dataset both mathematically and visualize it.
4. **Data visualization:** Visualizing the data to get better insight from it.
5. **Feature engineering:** We will also see what kind of stuff we can do in the feature engineering part.
6. Eliminating features that had an insignificant effect on the response variable by evaluating the p-values and R² value of the mode

**PROBLEM STATEMENT**:

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable.

**Model Building Phase**

After collecting/scraping the data, we have around 1948 rows and 9 columns. We need to build a machine learning model. Before model building, we will be doing data pre-processing steps. We will try different models with different hyperparameters and select the best model.

**ABOUT THE DATASET**

**About the data:**

1. Number of features in dataset: 9

2. Number of data points in dataset: 1948

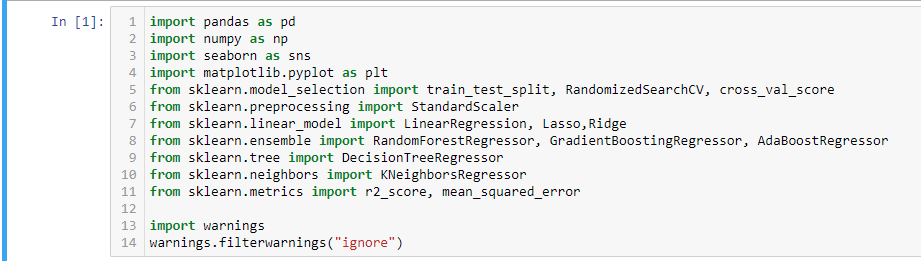
We have scraped price of 1948 rows from Yatra.com. This problem involves predicting the flight ticket prices of the old cars which are continuous and real-valued outputs. Thus, this is a **Regression Problem.**

**Features:**

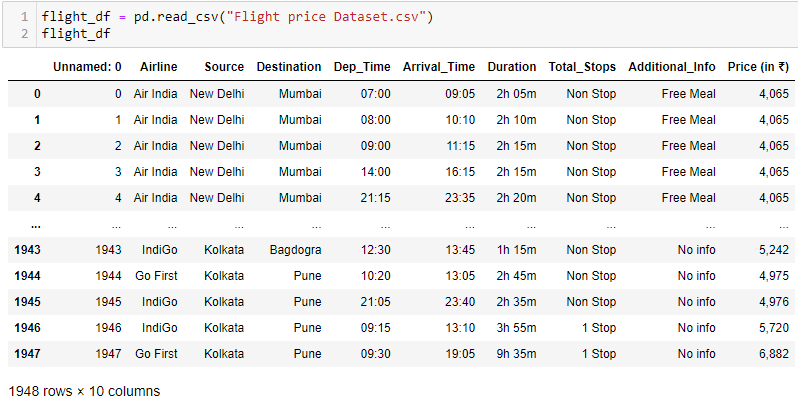
1. **Airline:** This column will have the names of all the types of airlines like Indigo, Jet Airways, Air India, and many more.
2. **Source:** This column holds the name of the place from where the passenger’s journey will start.
3. **Destination:** This column holds the name of the place to where passengers wanted to travel.
4. **Arrival\_Time:** Arrival time is when the passenger will reach his/her destination.
5. **Duration:**Duration is the whole period that a flight will take to complete its journey from source to destination.
6. **Total\_Stops:** This will let us know in how many places flights will stop there for the flight in the whole journey.
7. **Additional\_Info:** In this column, we will get information about food, kind of food, and other amenities.
8. **Price:** Price of the flight for a complete journey including all the expenses before onboarding.

**Importing Important Libraries:**

We need some libraries to be imported to work upon on dataset, we would import dataset by using pandas’ read\_csv method.



**Loading Data Set into variable: Here I am loading the dataset into the variable flight\_df .**

**.**

Dataset has been imported by using pandas read\_csv() function. We can see, it has mix of data types. Let’s check the shape of the dataset by calling shape method.

**Exploratory Data Analysis:**

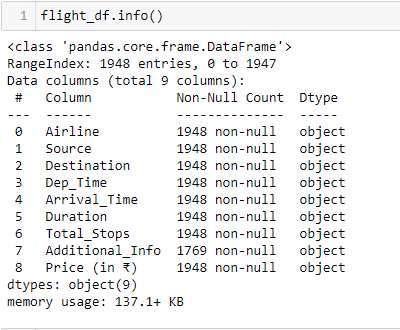
Before you start a machine learning project, it’s important to ensure that the data is ready for modelling work. Exploratory Data Analysis (EDA) ensures the readiness of the data for Machine Learning. In fact, EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. As we have two datasets, so we will do EDA for both the datasets simultaneously

Checking shape of the datasets**:**

****

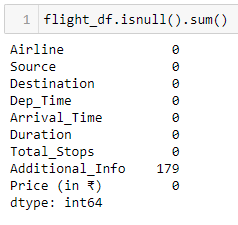
We can see there are 1948 rows and 9 columns.

**Getting detailed information about the datasets:**



Tha dataset has 1948 observations and 9 columns including target variable. Dataset all the variables as object data type. Target column is also object type .We will convert price column to int type.

**Checking the Missing Values**

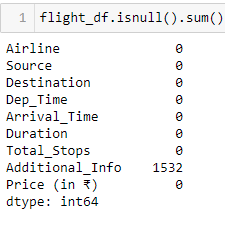


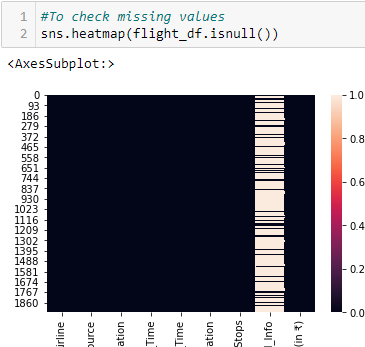
We can see that Additional info has 179 null values.

Also, there are many rows in Additional values which are having no info. So we will convert them also as nan value.



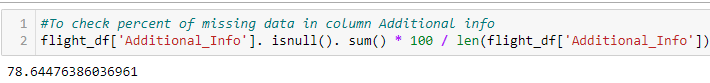
**Again, checking the null values**

****



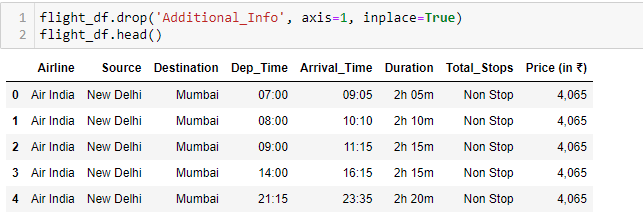
**Checking the Percent of null values :**

Both the null values were in the same row. So, by dropping the null value in axis=0(row wise), only one row is dropped.

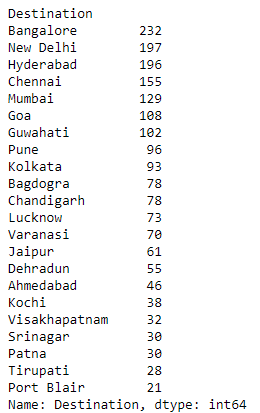
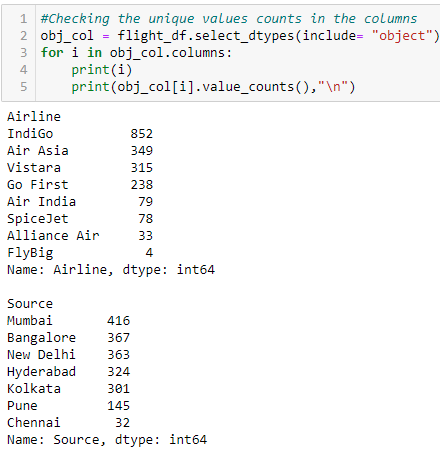


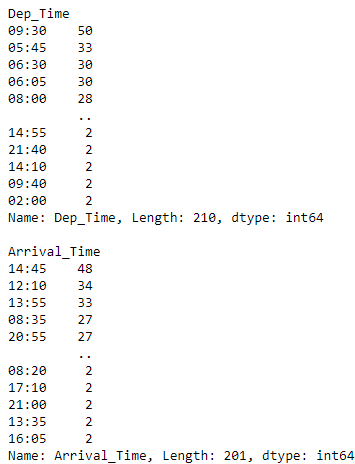
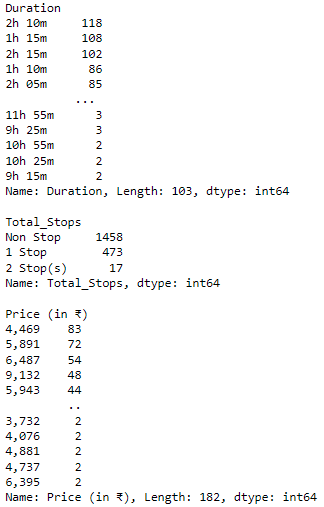
Here we can see more than 78% of data in Additional info is null. So we can drop the info column.

**Dropping the info column**

****

**Checking the unique values-counts of features in train data:**



**Conclusion:**

From the above value counts method, we have following conclusions:

1. we have multiple airlines data, top 3 airlines names are Indigo, AirAsia and Vistara.

2. Date column has to be converted into datetime columns and date and month from the date needs to be separated for analysis.

3. Major sources of the flights are from major 4 cities i.e. Mumbai, Bangalore, Delhi and Hydrabad.And their destination is also to major cities i.e. Bangalore, New Delhi, Hyderabad and Chennai.

4. Arrival time columns as multiple observations, it has hours, minutes

6. Duration is shown in hours and minutes.

7. Total stops tells that how many stops a flight takes. Most of the flights have no stop. Next to it are the flights which are having 1 stop.

### Creating features by separating Dep\_hour and Dep\_min from Departure Time and Arrival Time:

### we have created Dep\_hour and Dep\_min from column Dep\_time for better analysis and dropped the column Dep\_time.

### 

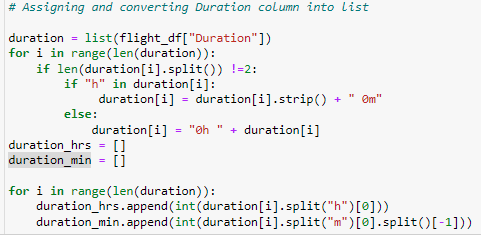
### 

Four new column Dep\_hour, Dep\_min, Arrival\_hour and Arrival\_min is created and Dep\_Time and Arrival\_Time is dropped.

### 

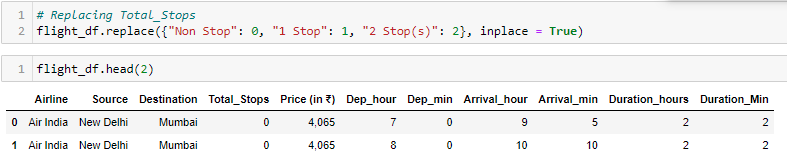
### Extracting the hours and min from the Duration column:

Since Duration column is showing Duration taken by a flight to cover the journey and it is showing in both Hour and min format. So we will first remove ‘h’ and ‘m’ from the column and separate the values in two separate columns as duration\_hrs and duration\_min .

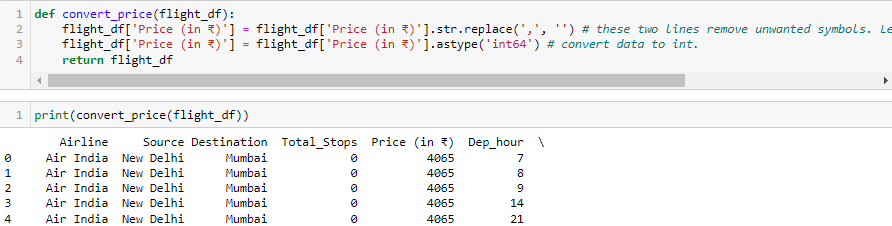




**Converting number for stops to numerical for easy analysis:**



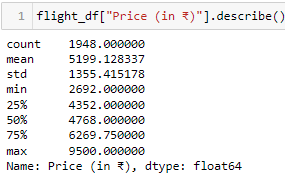
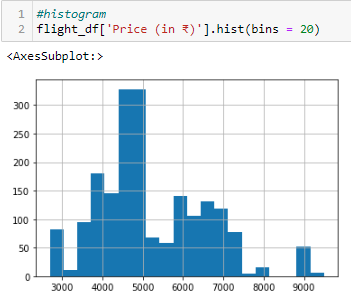
**Converting Target column (Price) to integer:**

****

**Univariate Analysis:**

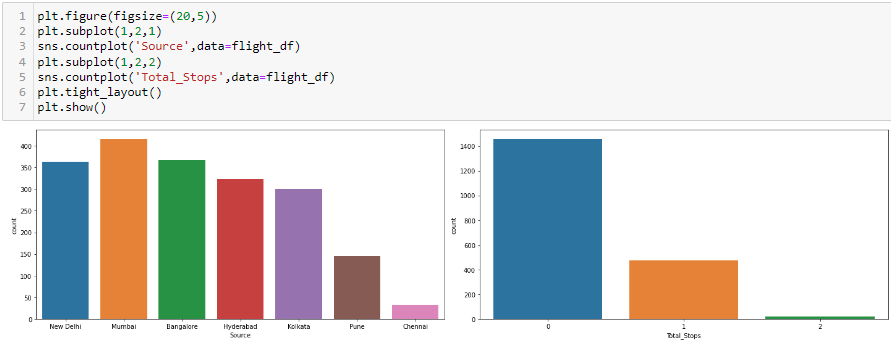
Uni means one, so in other words the data has only one variable. Univariate data requires to analyse each variable separately.  It doesn't deal with causes or relationships (unlike regression ) and it's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

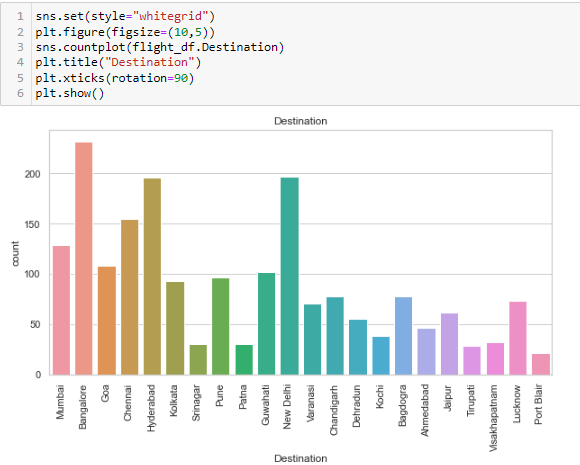
**Analysing Target Column from train data:**

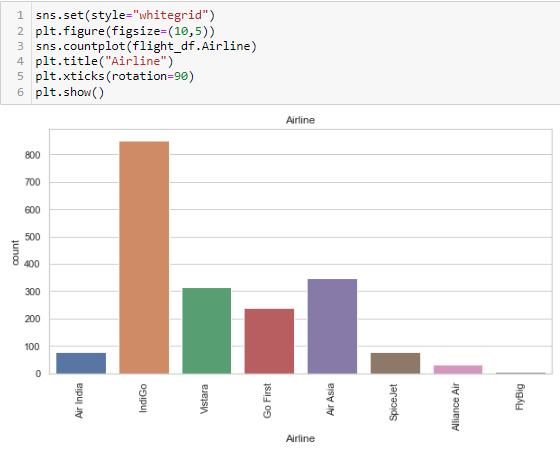
 

Price column has some outliers. Minimum Price is Rs 2692 and maximum price is Rs 9500.

**Handling Categorical Columns:**



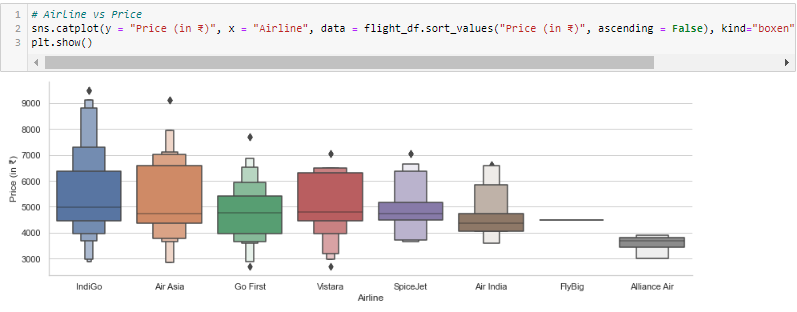




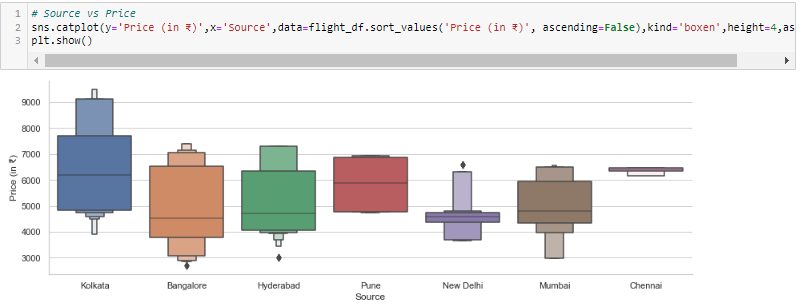
1. Most of the flight’s source is Mumbai, followed by Bangalore and maximum flight’s has destination as Bangalore followed by New Delhi and Hyderabad.
2. Maximum flights are having 0 stop only followed by one stop.
3. Indigo has the highest number of flights.

**Bivariate Analysis:**

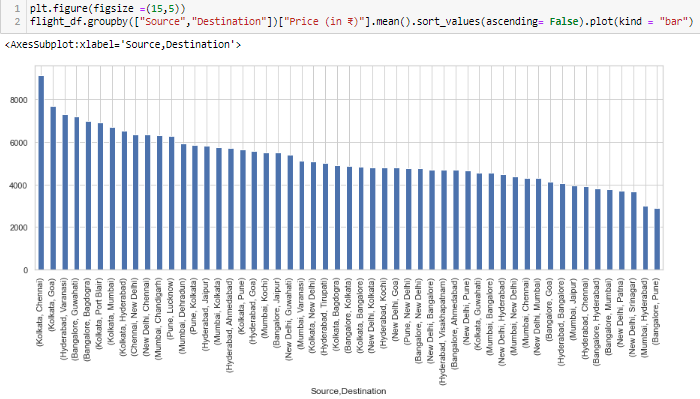
Bivariate analysis is finding some kind of empirical relationship between two variables. Specifically, the dependent vs independent Variables



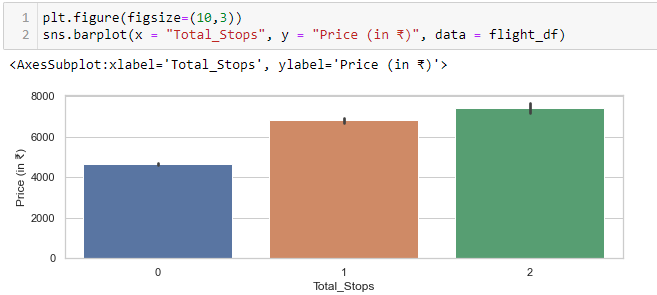
From graph we can see that Indigo has the highest Price.



Flights starting from Kolkata are having highest price and flight starting from Chennai are having lowest price.

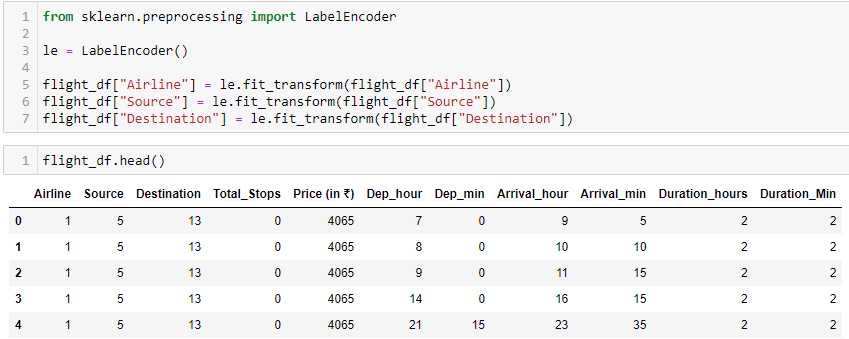


Kolkata to Chennai average price is Rs9500 approx., Bangalore to Pune average price is lowest which is around Rs 3000 approx.

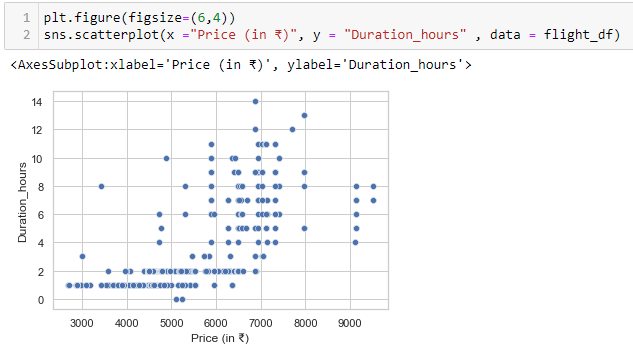


Here we can clearly see that wherever the number of stops is more, price is more. Price is highest for the flights having 2 stops.

**Converting Categorical data to Numerical Data using Label Encoder:**

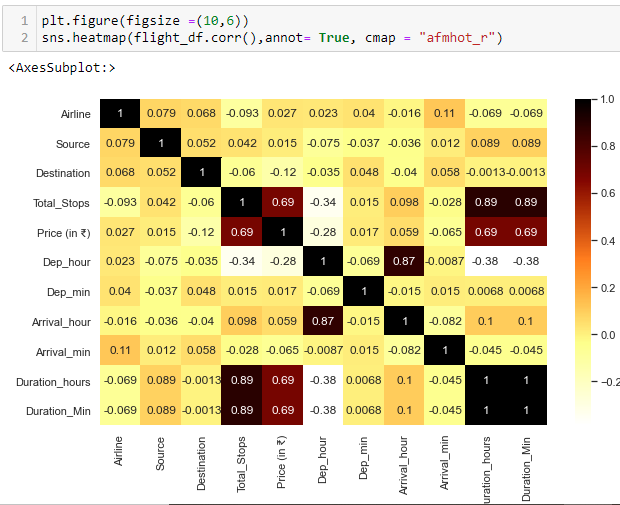


Now we can see all the columns are integer type.



More is the duration, price is less and vice versa.

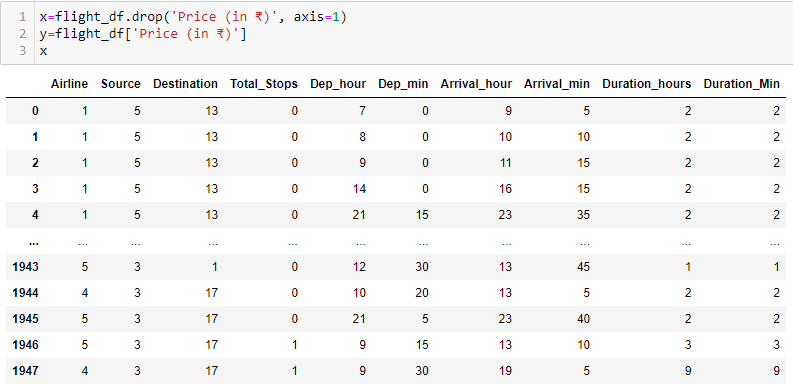
**Correlation Map:**



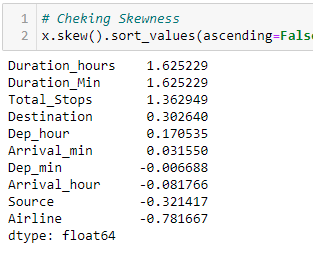
Conclusion:

Total\_stops ,Duration Minutes and Duration\_hours have positive correlation with Target column. Total\_stops and Duration hours are also correlation but we will keep the same in the dataset because there are only two which reflect maximum variance.

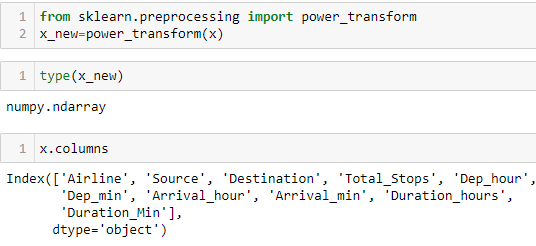
**Separating Independent and Dependent (target) features from Train Data:**

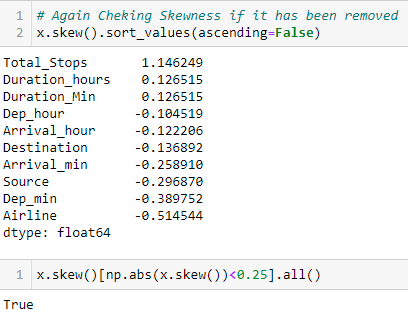


**Checking Skewness:**



Columns having skewness value less than -5 an greater than +5 are having skewed data. Here we can see Destination, Duration\_hours and Airlines have skewness. So we will apply power\_transform to remove the skewness.





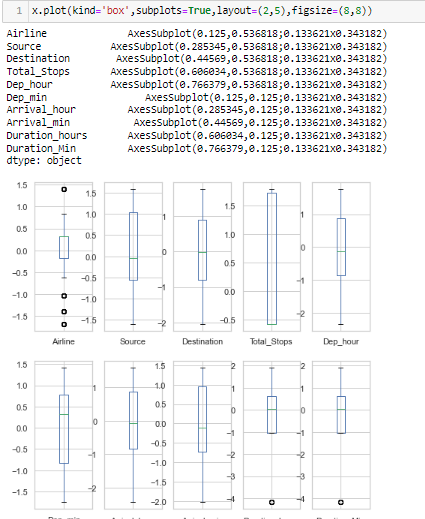
Here we can see now there is no skewness in any of the columns.

**Check for Outliers:**

#### Box Plot

#### This the visual representation of the depicting groups of numerical data through their quartiles. **Boxplot** is also used for detect the outlier in data set.

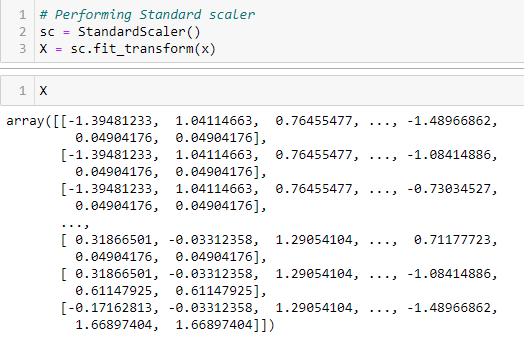
*I used box plot in this dataset because It captures the summary of the data efficiently with a simple box and whiskers and allows me to compare easily across groups.*



Here we can see there are not much Outliers in the dataset. So, we will not remove the outliers and proceed with the feature scaling.

**Features Scaling / Standard Scaler:**

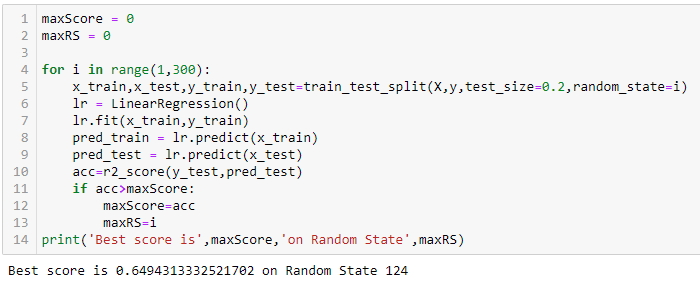
Feature Scaling is **a technique to standardize the independent features present in the data in a fixed range**. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.



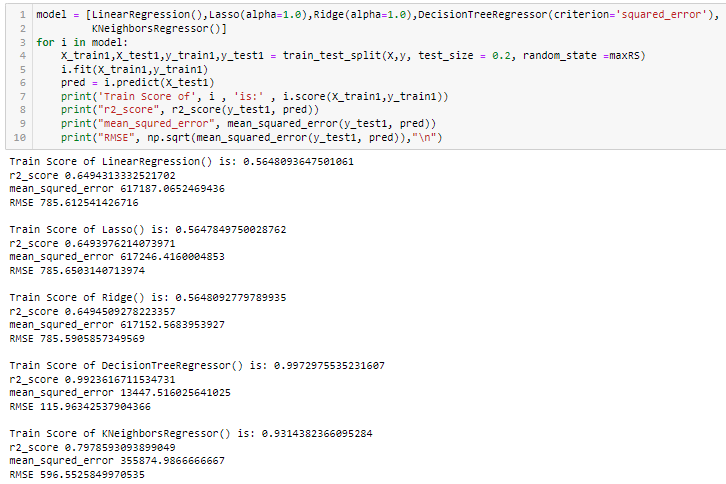
#### By using standard scaler, I have scaled the data in one range.

**Building Machine Learning Models:**

First I will find the best random state on which I will get the maximum score.



**Applying train-test split with Best Random State and applying ML on Different Algorithms:**



**Conclusions:**

Have checked Multiple Model and their score also. I have found that DecisionTreeRegressor() is working well on the dataset and have given less RMSE score . Now i will check with ensemble method to boost up score.

## Using Ensemble Technique to boost up score:

### RandomForestRegressor:

### 

### There is little difference between train score and test score. so, the model is overfitting.

### AdaBoostRegressor:

### 

### The difference between train and test score is least and RMSE is also very less. So selecting AdaBoost Model

### GradientBoostingRegressor:

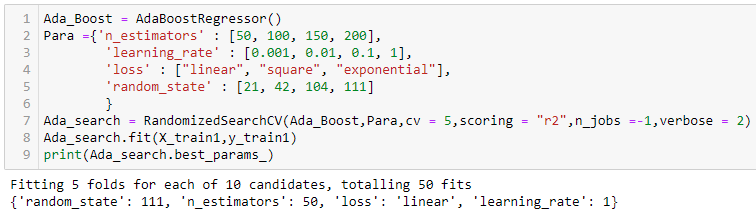
### 

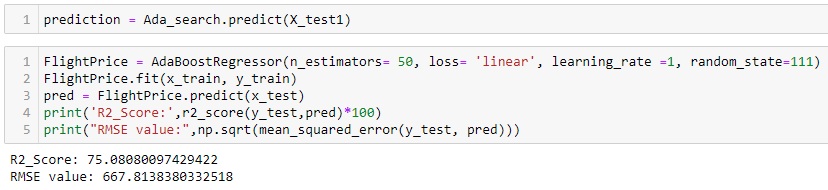
### **I have found that** **AdaBoostRegressor() is working well on the dataset with least train score and test score difference and have given less RMSE score . So i am selecting** AdaBoost**Regressor for final Model.**

## Hyper Parameter Tuning:

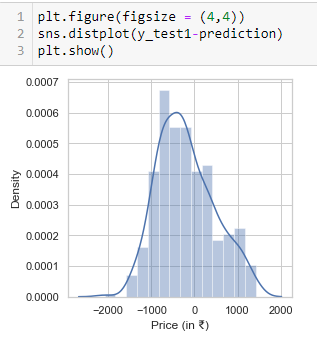
[**Hyperparameter tuning**](https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624) (or hyperparameter optimization) is the process of determining the right combination of hyperparameters that maximizes the model performance. It works by running multiple trials in a single training process.

We are using Randomsearchcv method for hyperparameter tuning to find best parameters for **AdaBoost**Regressor.



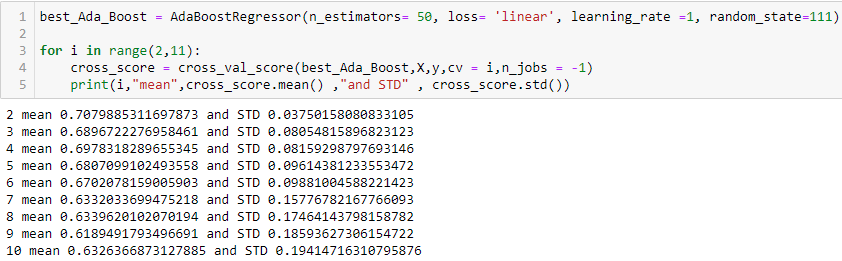


**The predicted y value is having a normalized curve which is good.**

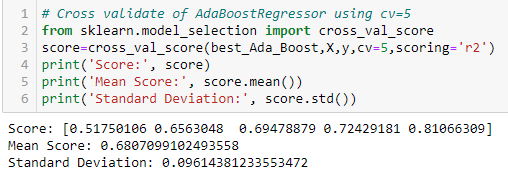
****

**Cross Validation:**

Cross-validation is **a resampling method that uses different portions of the data to test and train a model on different iterations**. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

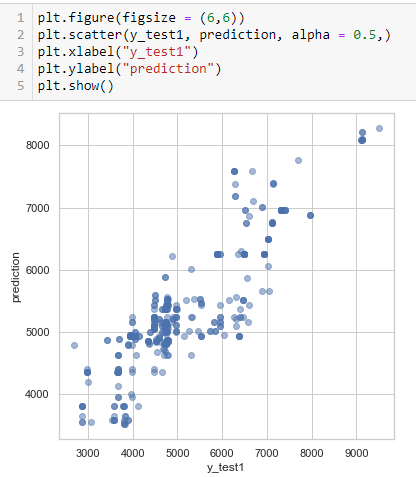


**Applying Cross validation Score=5**



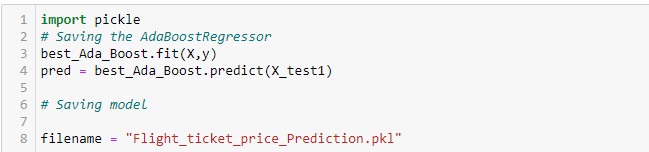
**Plotting y\_test1 vs predictions:**

* Simply plotting our predictions vs the true values.
* Ideally, it should be a straight line.



**Saving the Model:**

We are saving model by using python’s pickle library. It will be used further for the prediction.



**CONCLUSION:**

* After Scraping Flight Ticket prices for different source and destination cities like Delhi, Mumbai, Hyderabad, Bangalore, Chennai, Kolkata from different websites like Yatra.com, I have prepared an excel sheet and loaded the dataset for further EDA process.
* So, as we saw that we have done a complete EDA process, getting data insights, feature engineering, and data visualization as well so after all these steps one can go for the prediction using machine learning model-making steps.
* We have all the features categorical data types in the datasets and the dependent variable i.e. Price is also object data type.  **I am changing the target column to integer type and**  I applied the regression method for prediction.
* Once data has been cleaned and missing value is replaced, Label encoding is applied to them to convert them into Numerical ones. I trained the model on five different algorithms but for most of the models, train and test data was having a variance, and the model was overfitting.
* Only AdaBoost regressor worked well out of all the models, as there was less difference between train score and test score and RMSE was also low hence I used it as the final model and have done further processing.
* After applying hyperparameter tuning I got an accuracy(r2\_score) of 75% from the AdaBoostRegressor model after hyper parameter tuning which is a good score.

Then I saved the model.

**Limitations and Scope:**

* This study used only Yatra.com for web scraping. More websites can give more ideas and accurate reading. However, there was a relatively small dataset for making a strong inference because number of observations was only 1948. Gathering more data can yield more robust predictions.
* Secondly, there could be more features that can be good predictors. For example, here are some variables that might improve the model: Date of Journey, meal Details
* Another point that has room to improve is that the data cleaning process can be done more rigorously with the help of more technical information. For example, I had to drop meal info column because of lack of data..
* As a suggestion for further studies, while pre-processing data, instead of using a label encoder, one hot encoder method can be used. Thus, all non-numeric features can be converted to nominal data instead of ordinal data .

**I hope this article helped you to understand Data Analysis, Data Preparation, and Model building approaches in a much simpler way.**

**Thank you for reading this blog**