PROJECT REPORT

on

Flight Price Prediction project

By Sreekari I

INTRODUCTION

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 sale

DATA SOURCE

The data is collected from the https://www.makemytrip.com website. The details include the various details of the flights that are helpful to predict the price of the flight ticket.

OBJECTIVE

We are going to analyse the given data and check the factors that affect the price of the flight ticket and make a flight ticket price valuation model to predict the ticket price.

I have used jupyter notebook for data analysis.

DATA ANALYSIS

The given dataset has 7990 rows and 9 columns. The names of the columns are:

```
'Unnamed: 0',
'Name',
'Price',
'Departure time',
'Arrival time',
'Departure Place',
'Arrival Place',
'Duration',
'Flight type'
```

Now, let us analyse the clean and pre-process the data.

-→ Data Analysis

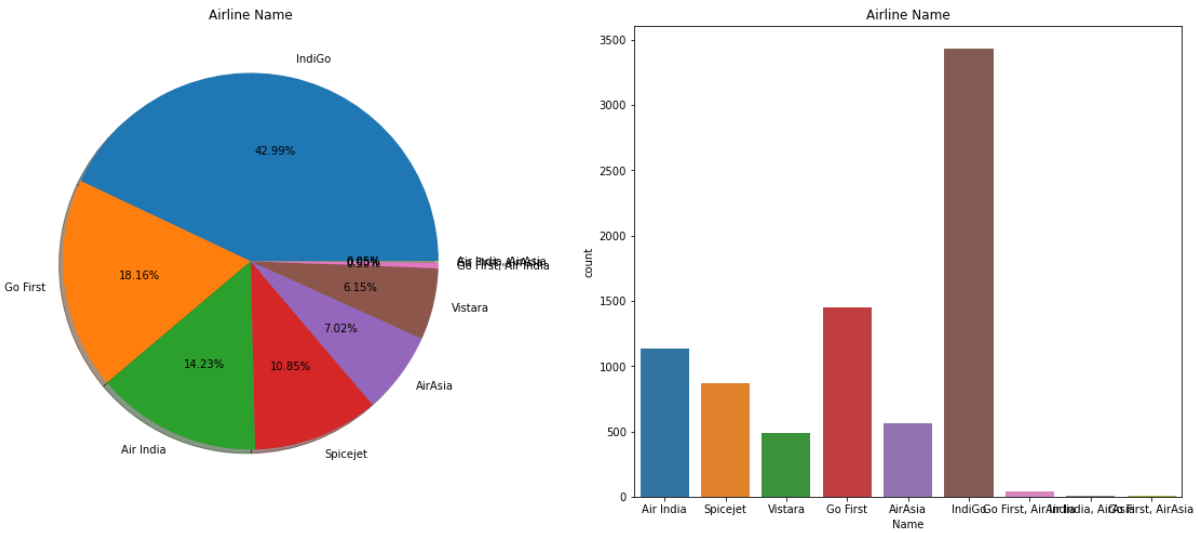
By checking the df.info(), I see that there are 8 columns of object datatype and one column of int64 datatype. Also we can see that there are no missing or null values.

After this, I have converted 'Departure time', 'Arrival time', 'Duration' to timestamps and these new columns that I have created were added to the dataset.

-→ Removing the columns that are not useful.

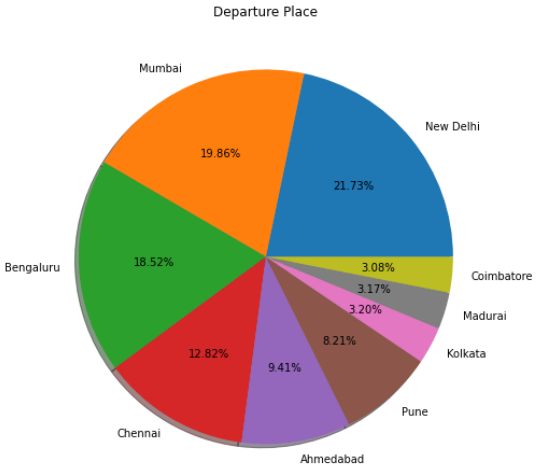
I have removed the 'Unnamed: 0' column since it is not useful in creating the model. Also I have dropped 'Departure time', 'Arrival time', 'Duration' columns as I have created timestamps for the required info from these columns

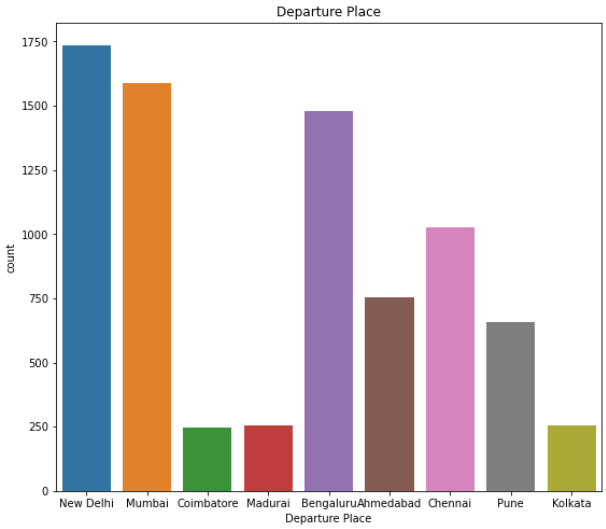
The below image is the pie chart and count plot showing the Airline Name column distribution



From the above image, we can see that there are ore flights of IndiGo Airlines followed by Go First Airlines while Air India, AirAsia being the least.

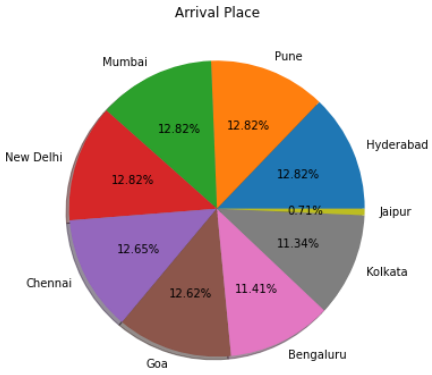
The below image is the pie chart and count plot showing the various Departure places of the flights

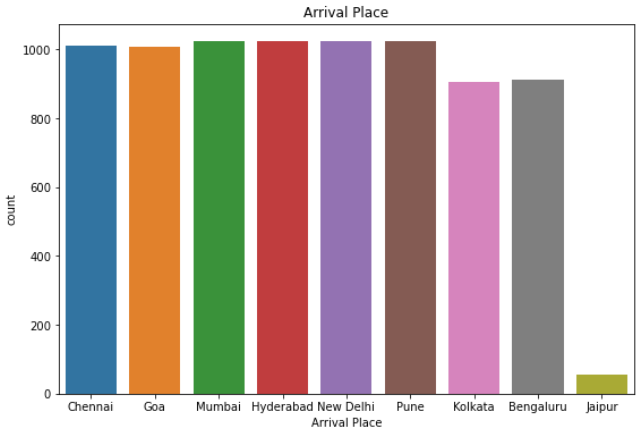




Here, we can see that most of the flights in the data have New Delhi as departure place while Coimbatore has least flights departured.

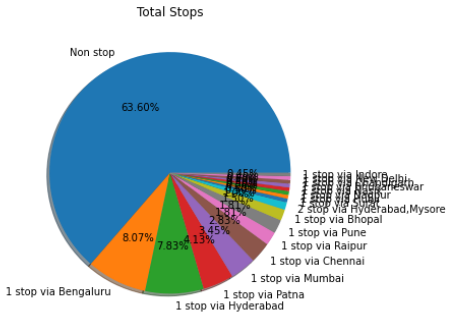
The below image is the pie chart and count plot showing the various Arrival places of the flights

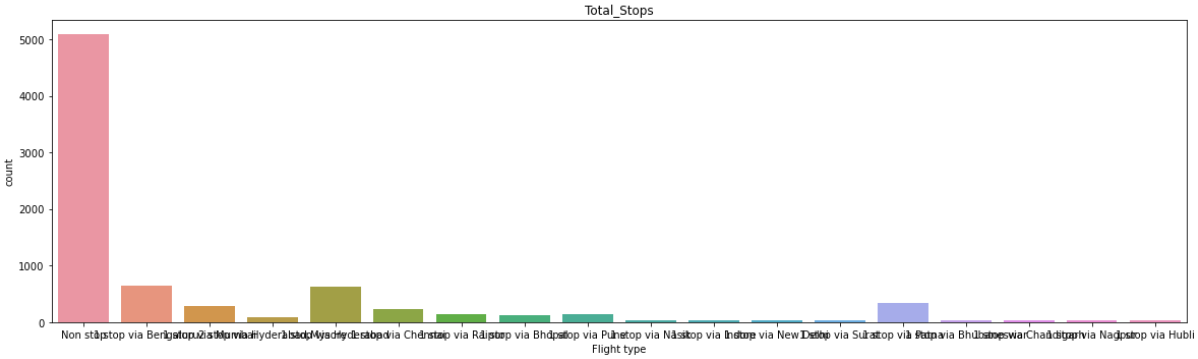




Here, we can see that most of the flights in the data have Hyderabad, Pune, Mumbai New Delhi as departure place while least flights were arrived at Jaipur

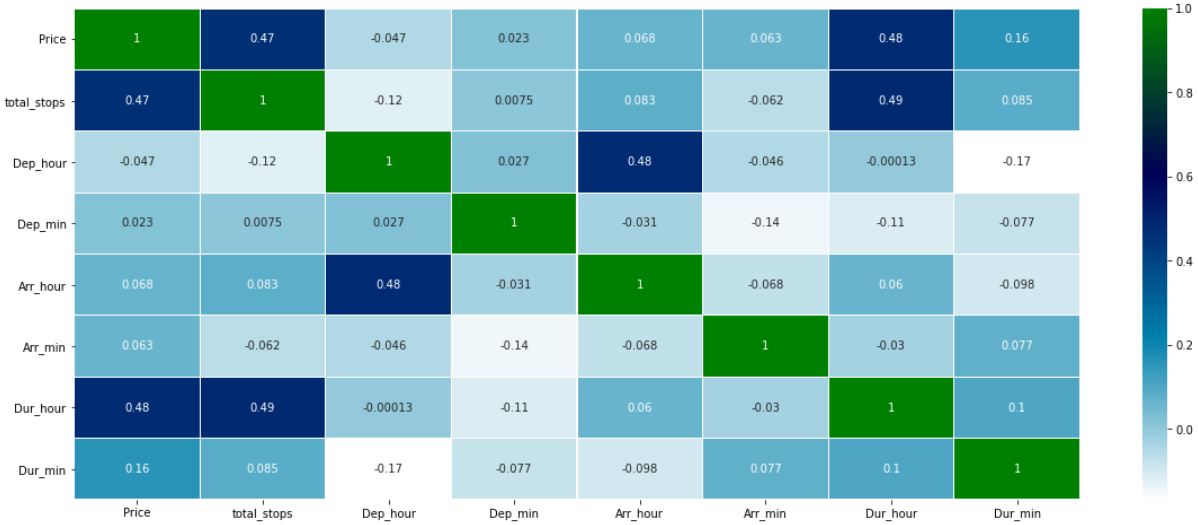
From the below images, we can see that most of the flights are Non stop.





I have encoded the categorical data to create data model. After all the above steps the dataset now have 7990 rows and 33 columns.

The below image is the heat map. We can see the correlation between the columns of the dataset with this heat map.



MODEL CREATION

Firstly, dividing the data into X and y sets for the model building.

```
X = df_train.loc[:, [ 'total_stops', 'Dep_hour', 'Dep_min', 'Arr_hour', 'Arr_min',
      'Dur_hour', 'Dur_min', 'Name_Air India, AirAsia', 'Name_AirAsia',
      'Name_Go First', 'Name_Go First, Air India', 'Name_Go First, AirAsia',
      'Name_IndiGo', 'Name_Spicejet', 'Name_Vistara', 'Bengaluru', 'Chennai',
      'Coimbatore', 'Kolkata', 'Madurai', 'Mumbai', 'New Delhi', 'Pune',
      'Chennai', 'Goa', 'Hyderabad', 'Jaipur', 'Kolkata', 'Mumbai',
      'New Delhi', 'Pune', 'Price']]
X.head()
```

```
y = df_train.iloc[:, 0]
y.head()
```

Next, splitting the data into training and test data test.

```
We specify this so that the train and test data set always have the same rows, respectively
```

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
random_state = 42)
```

I have used Linear Regression model and Random Forest Models for analysis.

LINEAR REGRESSION MODEL:

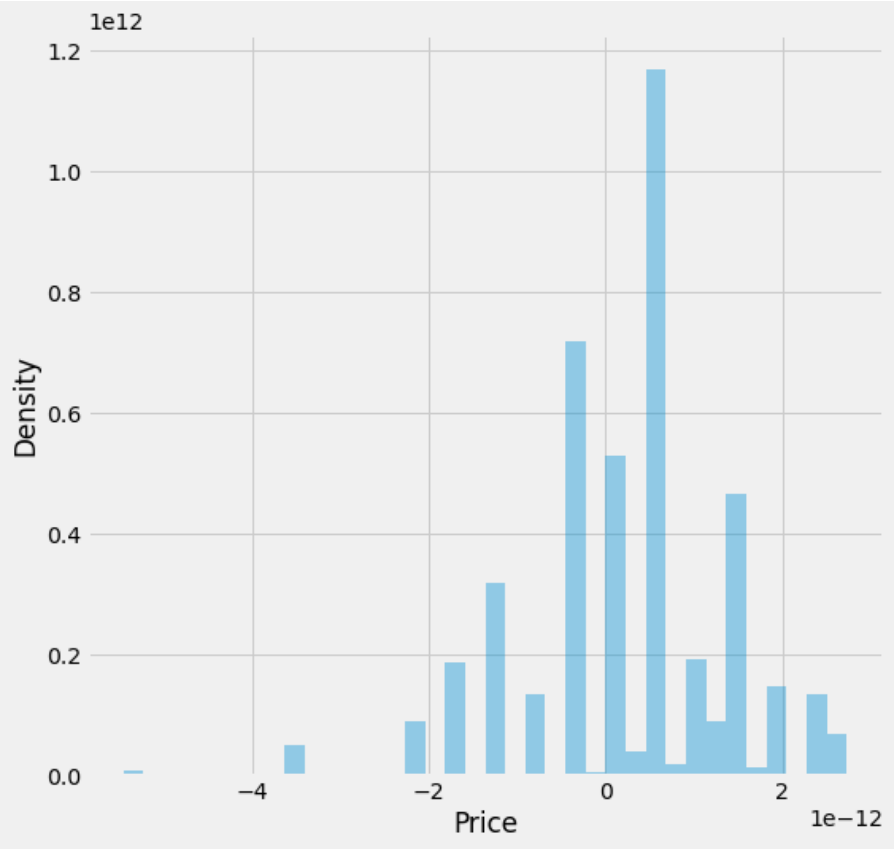
Accuracy is 96.68335419274092.

Regression Coefficients are [1.38042031e-13 -2.87547763e-14 -1.31733274e-15 5.388
48480e-15 1.15510899e-15 -1.41076387e-13 -6.64138883e-15 8.72941152e-13
 5.43390546e-13 1.97046867e-13 1.30657800e-12 2.75426519e-12
 1.59944059e-13 9.58558345e-14 1.14463895e-13 2.23245236e-12
 1.17671517e-12 -9.71661038e-14 2.11293631e-12 3.06491096e-14

-1.10734272e-12 2.72064946e-13 8.25284296e-13 -7.72162418e-14
 5.24916109e-13 -4.47453777e-13 8.02191328e-13 -2.43354802e-13
 1.17671775e-12 -9.71664155e-14 -1.16716231e-12 -4.86868168e-14
-5.24329512e-13 3.06489809e-14 -1.10734276e-12 8.25284336e-13
-7.72163366e-14 5.24916014e-13 -4.47453879e-13 8.02191131e-13
-2.43355290e-13 5.00000000e-01 5.00000000e-01]

Variance score: 1.0
Regression intercept : 160.20766244703452

MAE: 8.92704905964229e-13
MSE: 1.3635798051425615e-24
RMSE: 1.1677241990909331e-12



Prediction

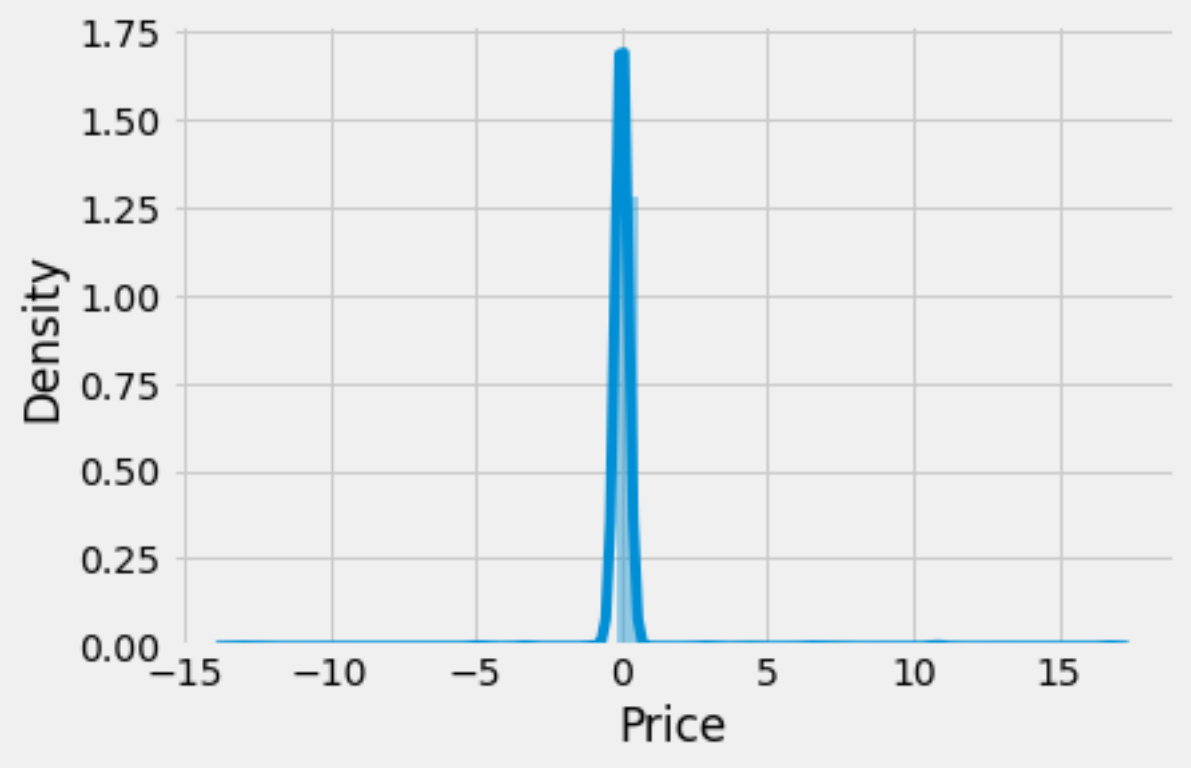
	Actual	Predicted
4953	2520	2520.0
5464	1715	1715.0
7487	2126	2126.0

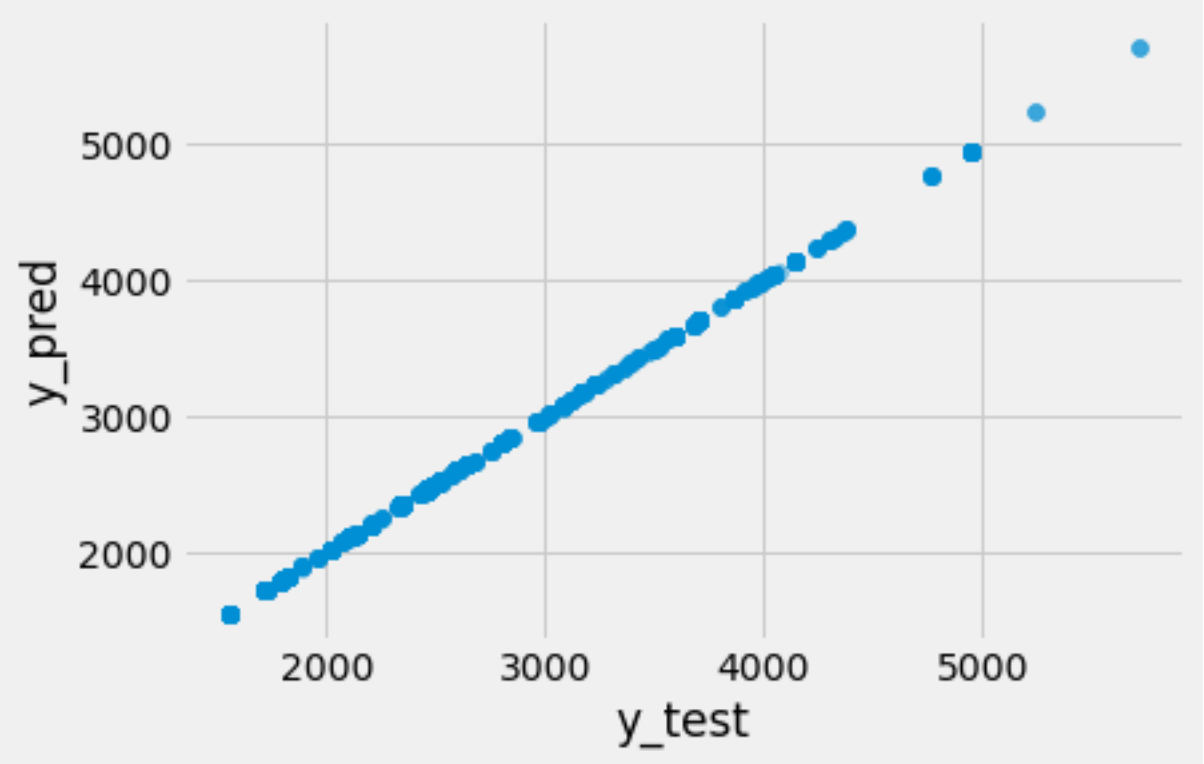
7608	2125	2125.0
2653	1830	1830.0
4623	3702	3702.0
1087	3988	3988.0
3837	2518	2518.0
4112	2463	2463.0
1783	4026	4026.0

RANDOM FOREST MODEL:

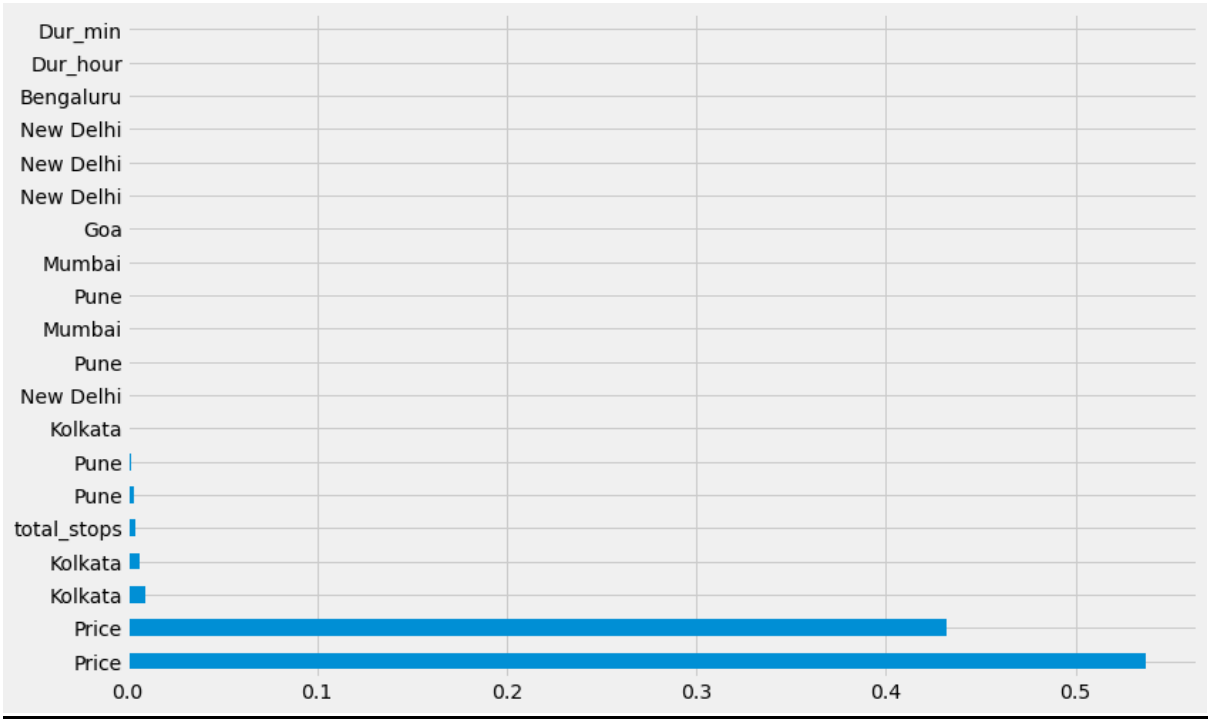
RandomForestRegressor Score: 0.9999998585639498

MAE: 0.08637672090112625
MSE: 0.9066258448060094
RMSE: 0.9521690211333329





Feature importances for better visualization



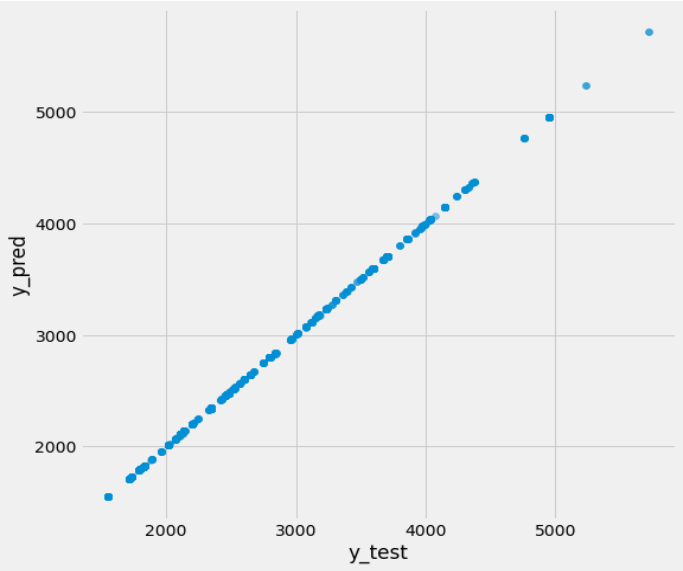
Hyper Parameter Tuning:

I have used RandomizedSearchCV for hyper parameter tuning. I have applied hyper parameter tuning for random forest model.

best_params_ :

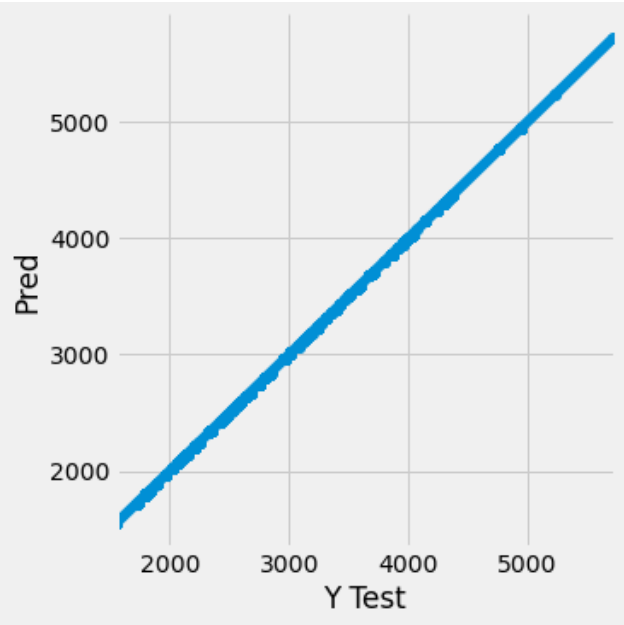
'n_estimators': 700,
 'min_samples_split': 15,
 'min_samples_leaf': 1,
 'max_features': 'auto',
 'max_depth': 20

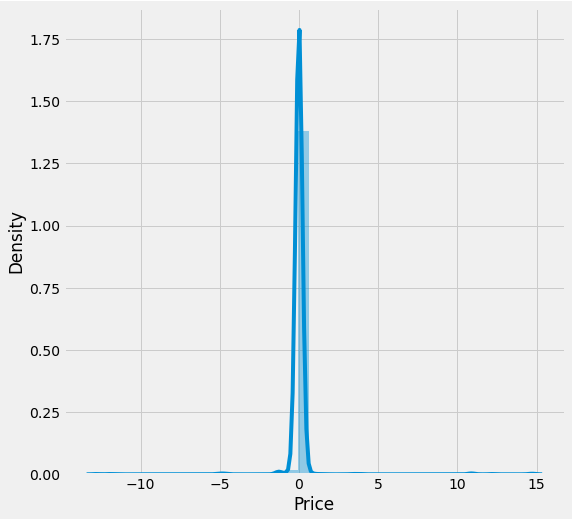
MAE: 0.0973402805559119
MSE: 0.8691632192061807
RMSE: 0.9322892358094567



PREDICTION

	Actual Data	Predicted Data
4953	2520	2520.0
5464	1715	1715.0
7487	2126	2126.0
7608	2125	2125.0
2653	1830	1830.0





CONCLUSION

From the above analysis we can see that various features of the flights affect the ticket price of the flights. If it has less stops and less duration of travel, the price of the ticket will be more when compared to the flight with more number of stops. The price will also be affected based on the distance of travel also. So the flight ticket price can be predicted based on these factors.