Flight Fare Prediction

Department of Computer Science
New Jersey Institute of Technology,
Newark, New Jersey - 07102

Shaily Sahay
ss4596@njit.edu

Table of Contents

Background

The ‘Flight Fare Prediction’ project builds a Machine Learning model using historical flight data that can accurately predict the cost of airline tickets.

This model could be useful for airlines and third-party travel sites to help them set competitive ticket prices, thereby giving them an edge over other players in the market. Its use case also extends to various travel agencies where they can analyze flight prices and offer attractive deals to their customers. For travelers, this information could help them find the best deals on flights and save money on their travel expenses.

Accurately predicting the cost of airline tickets can be a challenging task for several reasons, including the complex and dynamic nature of the airline industry, the wide range of factors that can influence ticket prices and the large amount of data that must be analyzed.

Scope of Study

The scope of the study is restricted to analysis of flight fares within Indian cities. It involves collecting and analyzing historical flight data from different airlines operating within India, including information about departure and arrival times, duration of flights, routes, ticket prices, and other relevant factors that determine price of tickets. The data used spreads over 4 months specific to the time period 03/01/2019 to 06/30/2019.

This is a public dataset and has been obtained from Kaggle, the link of which is given below:

Source of data: **https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh**

Data Description

CASE ⟶ flight ticket details for 10683 tickets of various airlines

VARIABLES ⟶ the dataset has 11 variables:

- **Airline** - It is a categorical variable that gives the name of the airline
- **Date_of_Journey** - It is a categorical variable that gives the date of departure
- **Source** - It is a categorical variable that gives the city of departure
- **Destination** - It is a categorical variable that gives the city of arrival

- **Route** - It is a categorical variable that gives the route that the flight will take to reach the destination. It lists all the cities where the flight will stop at.
- **Dep_Time** ⟶ It is a quantitative variable that gives the time of departure. The unit of measurement is 'hh: mm'
- **Arrival_Time** ⟶ It is a quantitative variable that gives the time of arrival. The unit of measurement is 'hh: mm'
- **Duration** ⟶ It is a quantitative variable that gives the total time the flight takes from source to destination. The unit of measurement is 'hh: mm'
- **Total_Stops** ⟶ It is a quantitative variable that gives the number of stops the flight takes to reach the destination. There is no unit of measurement
- **Additional_Info** ⟶ It is a categorical variable that provides information regarding services like baggage allowance, flight meals, class of ticket, etc.
- **Price** ⟶ It is a quantitative variable that gives the total price of the ticket. The unit of measurement is 'rupees'

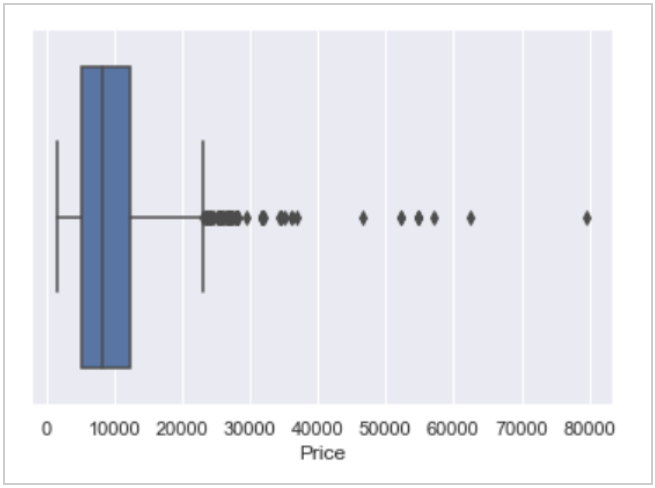
It must be noted that the Delhi city has 2 variations - 'Delhi' and 'New Delhi'. These two will be treated as different cities since they each comprise different areas.

Methodology

Data Pre-processing

Following pre-processing steps were taken:

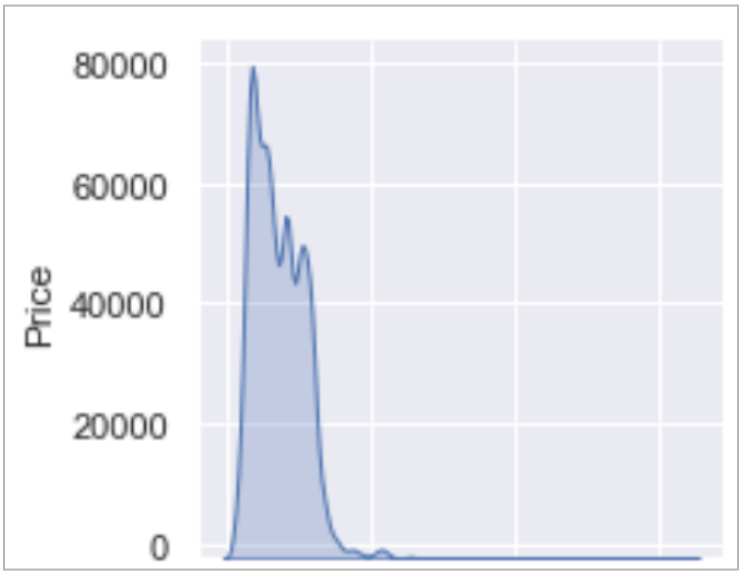
- The data was inspected for null values. There were 2 null values present, and those rows were dropped.
- The 'Duration' column was broken down to extract 'hours' and 'minutes' values. Separate columns were created to store these values: 'DurationHours' and 'DurationMinutes'
- The 'Date_of_Journey' column was broken down to extract 'day' and 'month' values. Separate columns were created to store these values: 'JourneyDay' and 'JourneyMonth'
- The 'Dep_Time' column was broken down to extract 'hours' and 'minutes' values. Separate columns were created to store these values: 'DepartureHour' and 'DepartureMinute'
- **Outliers**: Following 2 columns contained outliers:
    - 'Price' ⟶ These values indicate the cost of business class/ first class tickets, which are important to our investigation.



- ‘DurationHours’ → Some flights do have unusually long durations, and these flights offer low-cost tickets. Imputing these values would give an inaccurate result when training the model. These values should be left unchanged, as it would help the model learn the strength of negative correlation between 'DurationHours' and 'Price'.

Exploratory Data Analysis

- **Data Distribution**: Barring 2 columns ('Price' and 'DurationHours'), no other columns show high skewness. However, the distributions are multi-modal. 'Price' and 'Duration' columns are highly positively skewed.
    - 'Price' ⟶ Since this is a dependent variable, it was left unchanged



    - 'DurationHours' ⟶ A new column was created 'Duration_SQRT', which contained the Square root transformation of the 'Duration' column to treat skewness. The skewness reduced from 0.8512 to 0.2889.

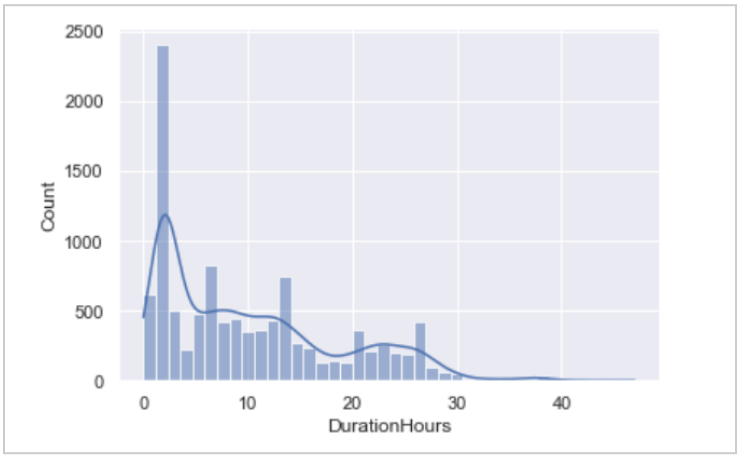


Fig 1: 'DurationHours' Original distribution

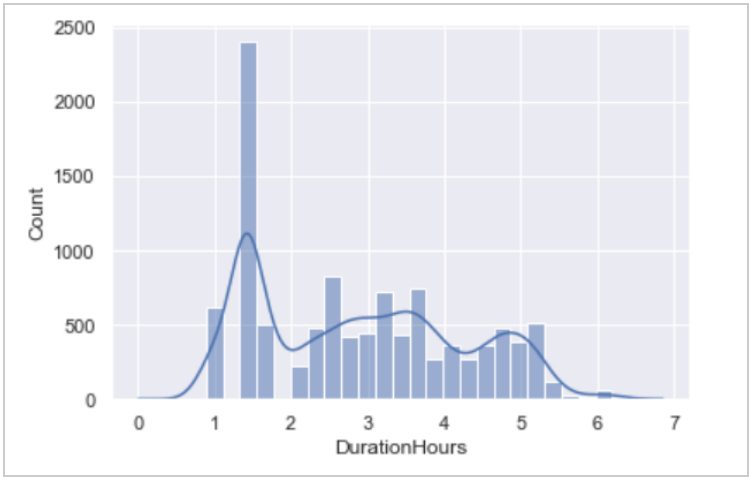
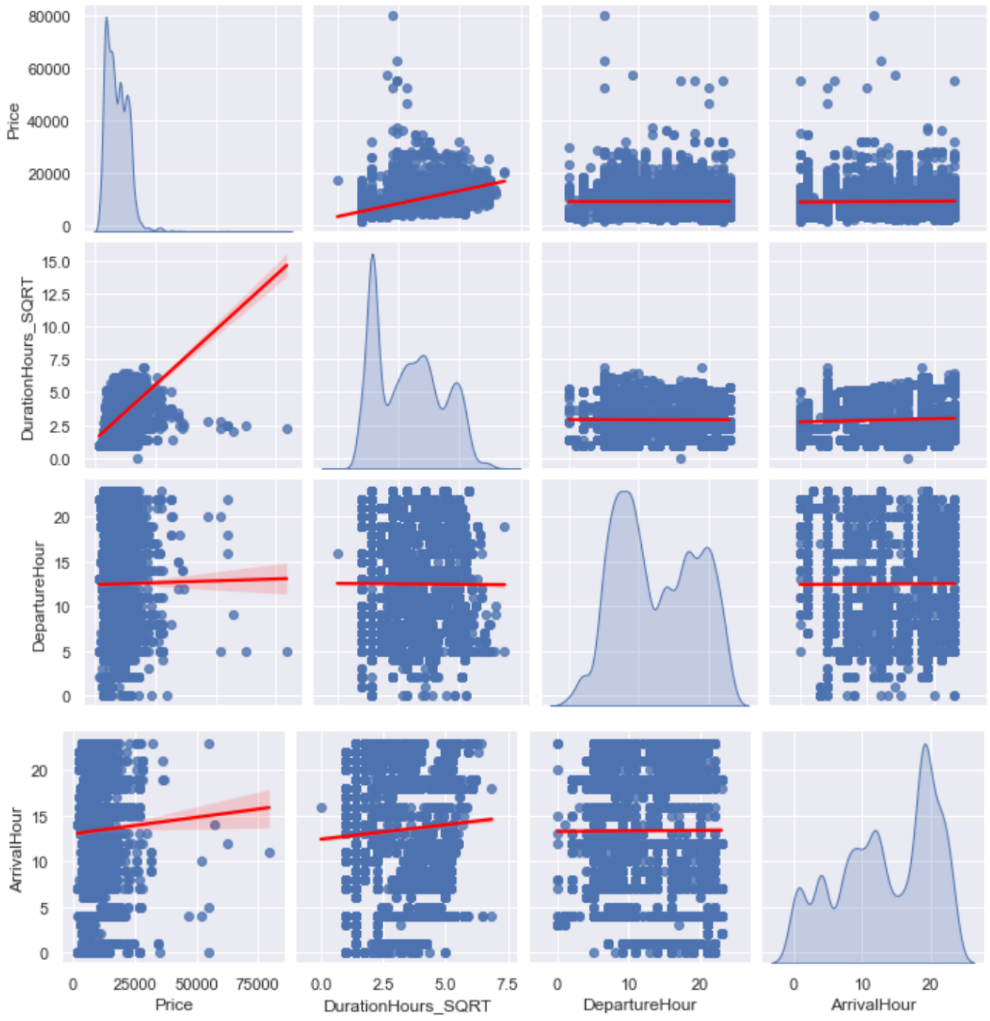


Fig 2: 'DurationHours' Transformed distribution

- **Relationship between 'Price' and numerical variables:** There is no evidence of a linear relationship between 'Price' and numerical independent variables. This can be confirmed from the below scatter plot. The points are very scattered around the regression line, indicating the absence of a linear relationship



- **Relationship between 'Price' and categorical variables:** Using cat-plots, the variation in price was observed with respect to 'Airlines', 'Source' and 'Destination' features.

  Airline vs Price

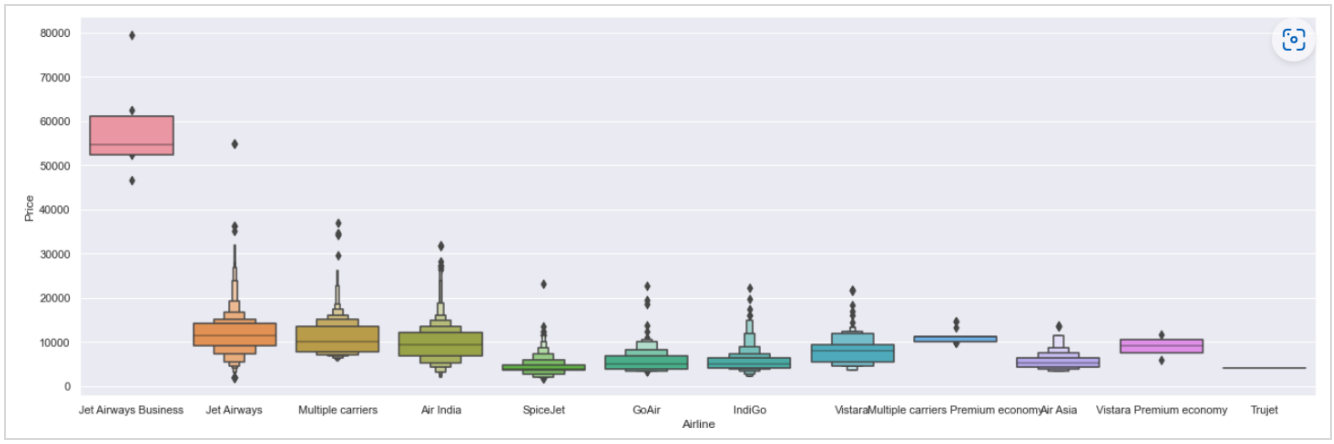


Fig 3: Price vs Airline

The graph shows that *Jet Airways Business* airline is the most expensive Airline with a mean price of approximately 50,500 rupees per ticket. We can also see that *Jet Airways, Air India* and *Vistara* have a similar per ticket mean price of 9000 to 12000 rupees, whereas *SpiceJet, GoAir, Indigo* and *Air Asia* offer tickets at almost half that price (approximately 5000 rupees) - the cheapest among all the carriers.

One common observation among all the carriers is the outlying price values. These values could indicate tickets sold during festivals and holidays, or last moment tickets which are generally sold at a very high price.

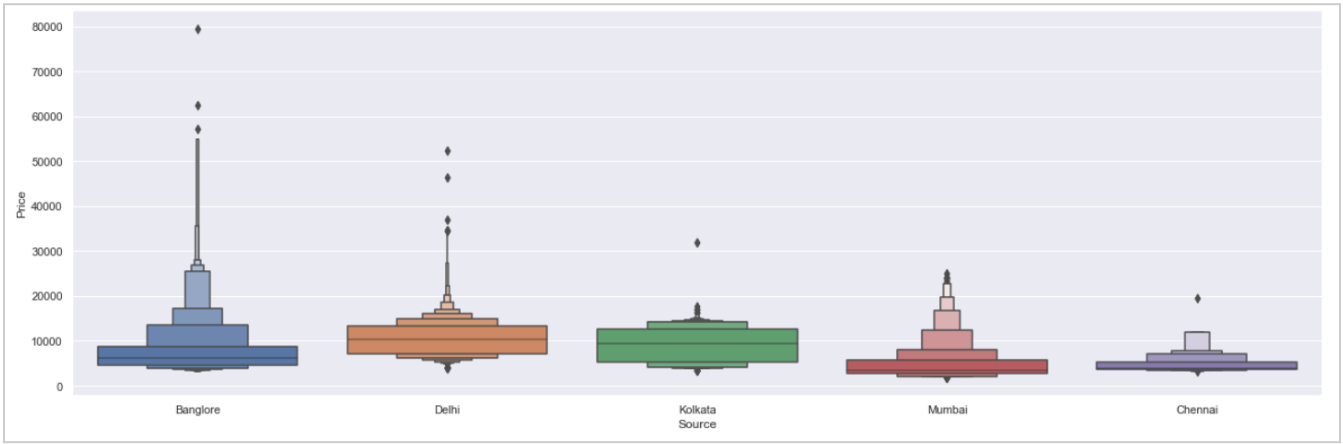


Fig 4: Price vs Source

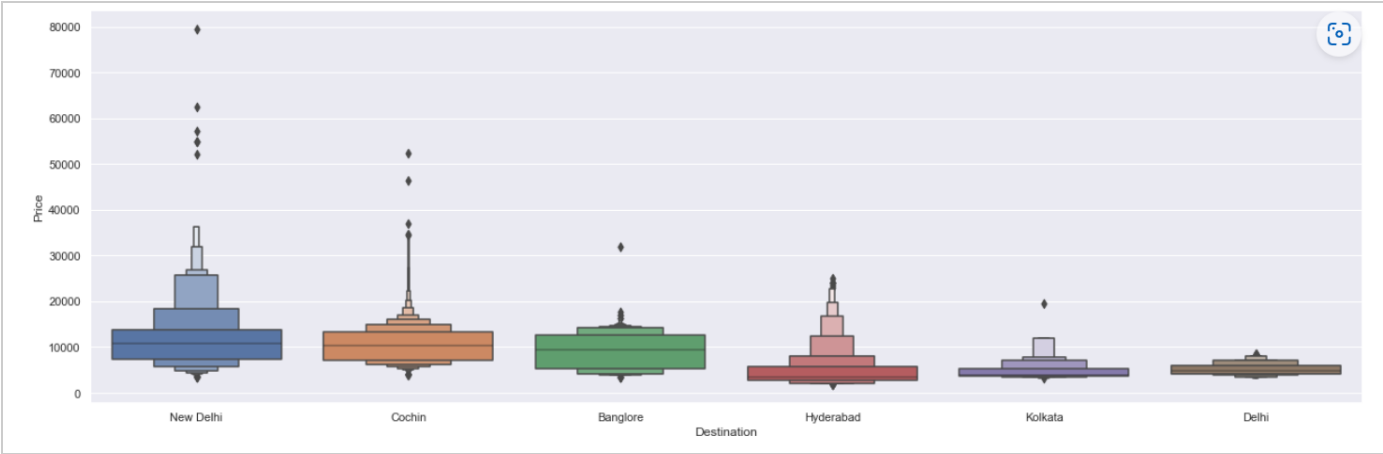


Fig 4: Price vs Destination

Some interesting observations were found when studying the effect of 'Source' and 'Destination' on 'Price'.

If one is flying *from* Kolkata or Delhi, the mean ticket price is very high (approximately 10000 rupees), but if they are going *to* Delhi, Kolkata, the mean price is cheapest. A similar observation was found for Bangalore. Flying *from* Bangalore is cheaper than flying *to* Bangalore.

Predictive Model

The next step in the process was to build a regression model which could predict flight prices based on the attributes present in the data. Following steps went into model training and prediction.

1. Encoding Categorical variables
2. Train-test split
3. Feature Scaling (Normalization)
4. Feature Selection

Encoding

Since Machine Learning models only understand numerical data, we need to convert categorical data into numerical values.

One-hot encoding was done on Nominal attributes - 'Airline', 'Source', 'Destination'. **Label encoding** was done on Ordinal attributes - 'Total_Stops'.

After encoding, the data frame had following values:

```
df_encoded.head()
```

	Total_Stops	Price	DurationMinutes	JourneyDay	JourneyMonth	DepartureHour	DepartureMinute	ArrivalHour	ArrivalMinute	DurationHours_SQRT
0	0	3897	50	24	3	22	20	1	10	1.414214
1	2	7662	25	1	5	5	50	13	15	2.645751
2	2	13882	0	9	6	9	25	4	25	4.358899
3	1	6218	25	12	5	18	5	23	30	2.236068
4	1	13302	45	1	3	16	50	21	35	2.000000

Airline_Air India	Airline_GoAir	Airline_IndiGo	Airline_Jet Airways	Airline_Jet Airways Business	Airline_Multiple carriers	Airline_Multiple carriers Premium economy	Airline_SpiceJet	Airline_Trujet	Airline_Vistara	Airline_Vis Prem econ
0	0	1	0	0	0	0	0	0	0	
1	0	0	0	0	0	0	0	0	0	
0	0	0	1	0	0	0	0	0	0	
0	0	1	0	0	0	0	0	0	0	
0	0	1	0	0	0	0	0	0	0	

Source_Chennai	Source_Delhi	Source_Kolkata	Source_Mumbai	Destination_Cochin	Destination_Delhi	Destination_Hyderabad	Destination_Kolkata
0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	0	0	1	0	0	0
0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0

Train-Test split

The train-to-test split ratio was 4:1.

Feature Scaling

Feature scaling is required for Linear regression models, but not for non-linear models like 'Random Forest'. Since we had to test multiple models to see which one fits best for the data, we used scaled features on Linear regression, and unscaled features on other non-linear models.

Sci-kit learn's *MinMaxScalar* function was used to do feature scaling.

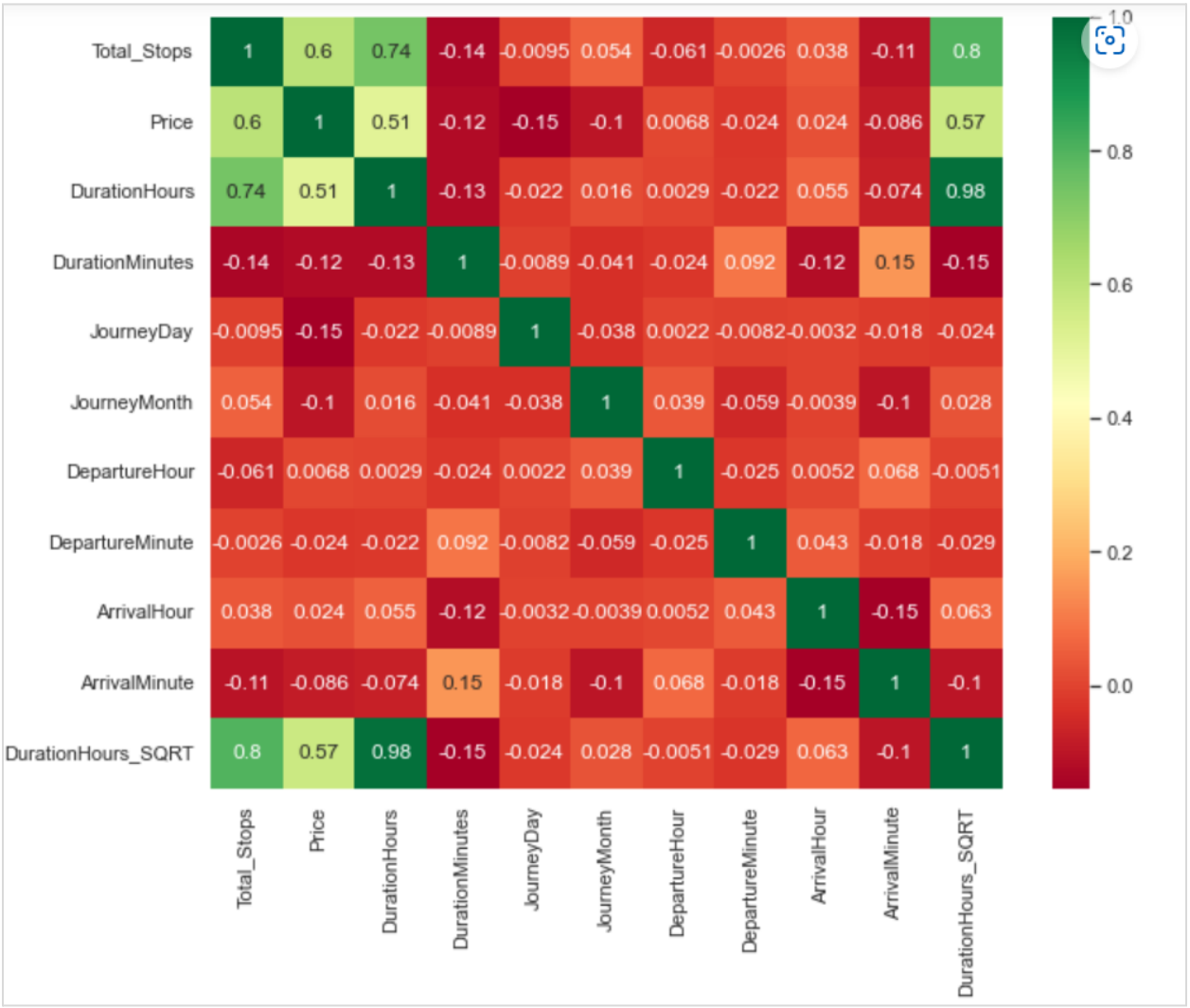
Feature selection

Identifying the best attribute that will aid and have a defining relationship with the target variable Here are a few techniques that were used to identify the most important features.

Heatmap

From the below heatmap, we can conclude that 'Price' is highly correlated to 'DurationHours_SQRT' and 'TotalStops'.

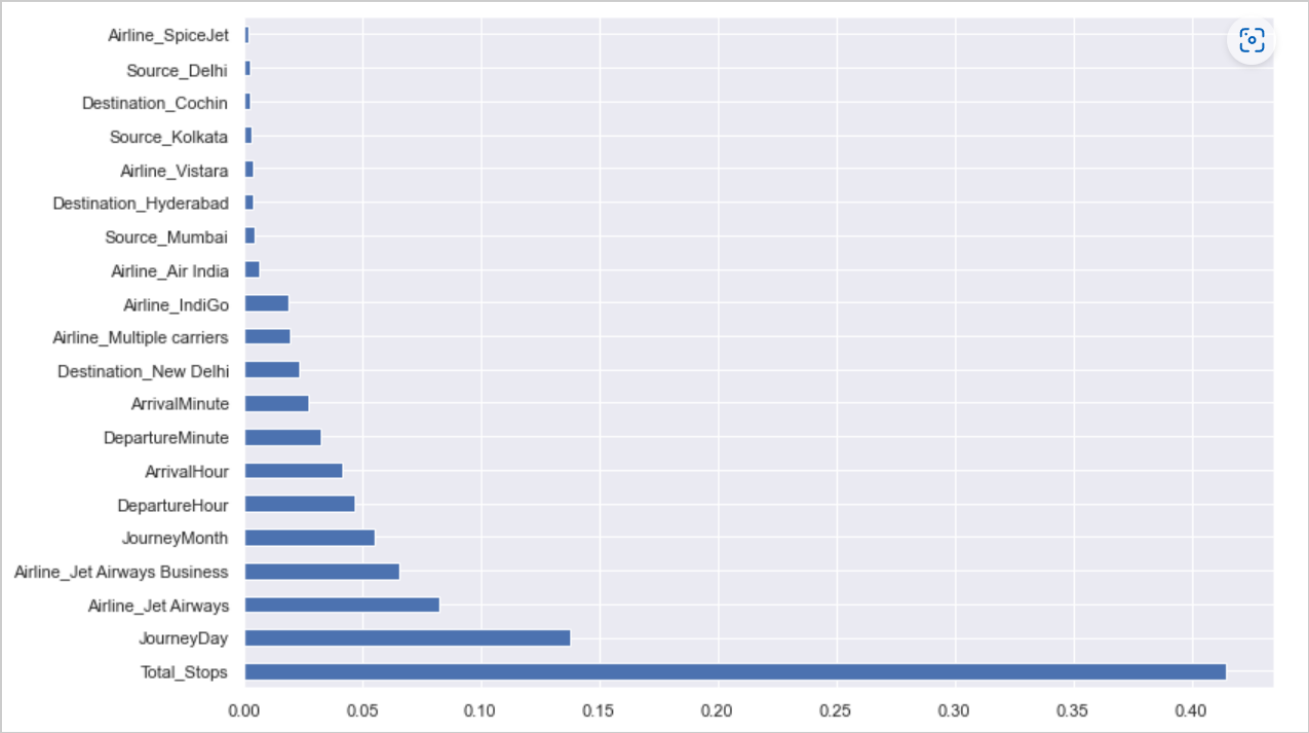
Multicollinearity was also observed between 'TotalStops' and 'DurationHours_SQRT' with a 80% correlation. So, 'DurationHours_SQRT' column was dropped to avoid multicollinearity.



Feature_importance

Once the model with highest accuracy was selected (discussed in the next section), the **feature_importance_** function of the model was used to obtain the feature importance score, which demonstrates how significant each feature is when predicting the value of the response variable.

The below graph shows that the 'Total_Stops' and 'JourneyDay' played the most significant role when predicting the flight fare. Similarly, importance of all other features can be observed from the graph.



Model Selection

As observed previously, there is no linear relationship between dependent and independent variables. Due to this the Linear Regression model did not perform well in this study. Following metrics confirm the finding:

```
--------- Metrics --------------
Training accuracy: 0.6321534562852504
Test accuracy: 0.5612190962781192
MAE: 2095.3528310715956
MSE: 10295193.116840197
RMSE: 3208.612335081974
```

Both train and test accuracy values are very low, and all the loss function values are high.

Next, non-linear models were tested. First Extra-tree regressor was used, which gave the below metrics:

```
--------- Metrics --------------
Training accuracy: 0.9569676971163795
Test accuracy: 0.80793646484062
MAE: 1236.4029049000185
MSE: 4506420.330503184
RMSE: 2122.8330905898333
```

As expected, the accuracy values have greatly increased and loss function values have decreased.

Random Forest Regressor was also tested to see if it provided better results.

```
--------- Metrics --------------
Training accuracy: 0.9569676971163795
Test accuracy: 0.80793646484062
MAE: 1236.4029049000185
MSE: 4506420.330503184
RMSE: 2122.8330905898333
```

As we can see it provides an 80% test accuracy, which is the highest among all the models.

The validity of the model's inference can be confirmed by the distribution of residual plots.

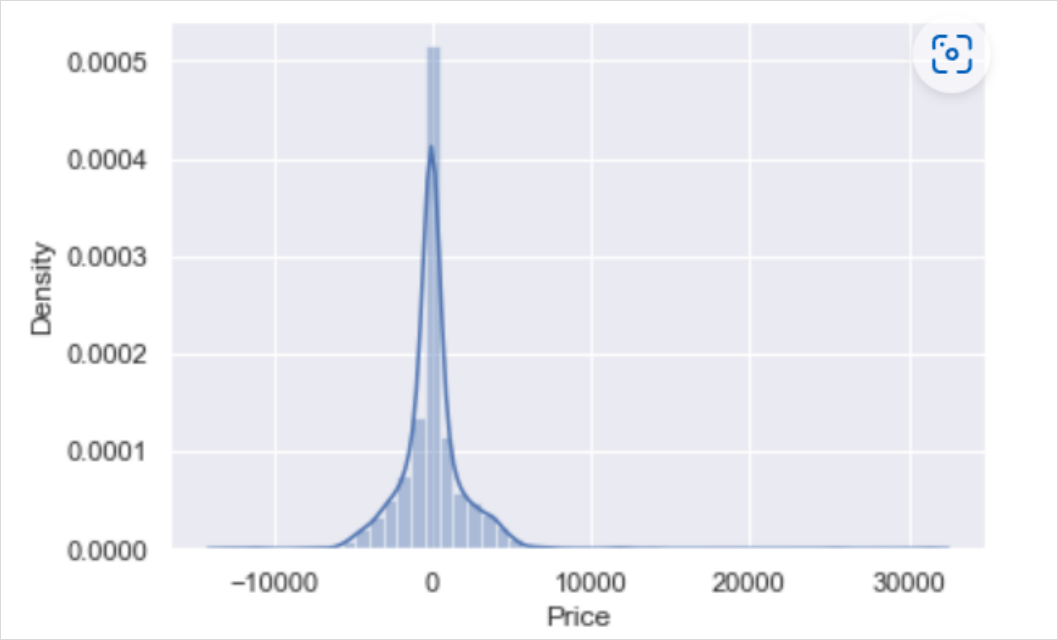


Fig 4: Distplot for test and prediction

Conclusion

The results of measurement tools (metrics) like MSE, MAE, RMSE, and R^2 effectively demonstrate that this algorithm is capable of reliably estimating the cost of a flight, which will enable the company to enhance the market value based on massive datasets in the future.

References

1. Tziridis, K., Kalampokas, T., Papakostas, G.A. and Diamantaras, K.I., 2017, August. Airfare prices prediction using machine learning techniques. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1036-1039). IEEE.
2. Biswas, P., Chakraborty, R., Mallik, T., Uddin, S.I., Saha, S., Das, P. and Mitra, S., 2022. Flight price prediction: a case study. *Int. J. Res. Appl. Sci. Eng. Technol.(IJRASET)*, *10*(6).
3. Subramanian, R.R., Murali, M.S., Deepak, B., Deepak, P., Reddy, H.N. and Sudharsan, R.R., 2022, January. Airline Fare Prediction Using Machine Learning Algorithms. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 877-884). IEEE.

4. Liu, T., Cao, J., Tan, Y. and Xiao, Q., 2017, December. ACER: An adaptive context-aware ensemble regression model for airfare price prediction. In *2017 International Conference on Progress in Informatics and Computing (PIC)* (pp. 312-317). IEEE.