FLIGHT PRICE PREDICTION PROJECT

Submitted by:

ANJAY KRISHNA T U

ACKNOWLEDGMENT

The data which need for this project is collected from the websites yatra.com and makemytrip.com. These collected information about flights are used for building a machine learning model for predicting the flight price.

INTRODUCTION

- Business Problem Framing

  The price of flight ticket varies over time, that may up or down but when the time of departure is near the price is always higher. Here we need to collect details of flight fares and other details and build a machine learning model for predicting the price of a flight ticket.

- Motivation for the Problem Undertaken

  The flight ticket fare is the one thing that varies with time, that may go up or down. This project may help to find out the reason the price varies over time and help customer for booking the flight tickets.

Analytical Problem Framing

- Data Sources and their formats

  The data for this project is collected from the web sites yatra.com and makemytrip.com. The collected data contains the information about the fights like Airline, Date of Journey, Source, Destination, etc. There are 1670 rows and 10 columns in the collected dataset.

- Data Preprocessing Done

  In the collected data set null values are present they are removed from the collection. All the features are showing object data types. In the source and destination Bangalore is showing two time that is replaced by correct one. The total stops are showing differently no stops, 1stops, 2stops, 3 stops are replaced by 0,1,2 and 3 respectively. Date of journey, Arrival time and Departure time are converted to date time format and new feature columns are created. From Duration new features like Duration Hour and Duration minute is created. The output variable price contains a symbol and comma they are removed from the price column. There is no skewness in the dataset. Outliers are present in Price and Duration Hour, they are removed using z-score method.

- Data Inputs- Logic- Output Relationships

  The input variables are Airline, Date of Journey, Source, Destination, Departure time, Arrival time, Duration and Total stops. The output variable is the Price. Price is positively correlated to the Total stops and Duration hour.

- Hardware and Software Requirements and Tools Used

  Selenium Webdriver – For scraping the data
  Anaconda jupyter Notebook – For Programming
  Pandas, Seaborn, Matplotlib – For Visualization

Model/s Development and Evaluation

- Testing of Identified Approaches (Algorithms)
    1. Linear Regressor
    2. Lasso Regressor
    3. Ridge Regressor
    4. Decision Tree Regressor
    5. Random Forest Regressor
    6. Ada boosting Regressor
    7. Gradient boosting Regressor
- Run and Evaluate selected models

  After training the data 3 algorithms are performing better.

  1. Decision Tree Regressor

```
dt=DecisionTreeRegressor(splitter='best',
                         min_samples_split=15,
                         max_features='auto',
                         max_depth=19,
                         criterion='friedman_mse')
dt.fit(x_train,y_train)
preddt=dt.predict(x_test)
print('R2 score :',r2_score(y_test,preddt))
print('Mean squared error :',mean_squared_error(y_test,preddt))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,preddt)))
```

```
R2 score : 0.7126295664983515
Mean squared error : 1305544.9031766327
Root Mean squared error  : 1142.6044386298493
```

  This ML algorithm gives R2 score of 71.26%.

  2. Random Forest Regressor

```
rf=RandomForestRegressor(min_samples_split=10,
                         min_samples_leaf=5,
                         max_features='auto',
                         max_depth=18,
                         criterion='poisson')
rf.fit(x_train,y_train)
predrf=rf.predict(x_test)
print('R2 score :',r2_score(y_test,predrf))
print('Mean squared error :',mean_squared_error(y_test,predrf))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predrf)))
```

```
R2 score : 0.3760767155162268
Mean squared error : 2834529.127111267
Root Mean squared error  : 1683.605989271619
```

The Random forest is a supervised machine learning algorithm uses and ensemble leaning method for regression problem. The R2 score of this model is 37.60%.

3. Gradient boosting regressor

```
gb=GradientBoostingRegressor(max_features='auto',
                             loss='huber',
                             learning_rate=0.1,
                             criterion='mse')
gb.fit(x_train,y_train)
predgb=gb.predict(x_test)
print('R2 score :',r2_score(y_test,predgb))
print('Mean squared error :',mean_squared_error(y_test,predgb))
print('Root Mean squared error  :',np.sqrt(mean_squared_error(y_test,predgb)))
```

```
R2 score : 0.6778033096950682
Mean squared error : 1463763.1356935352
Root Mean squared error  : 1209.8607918655498
```

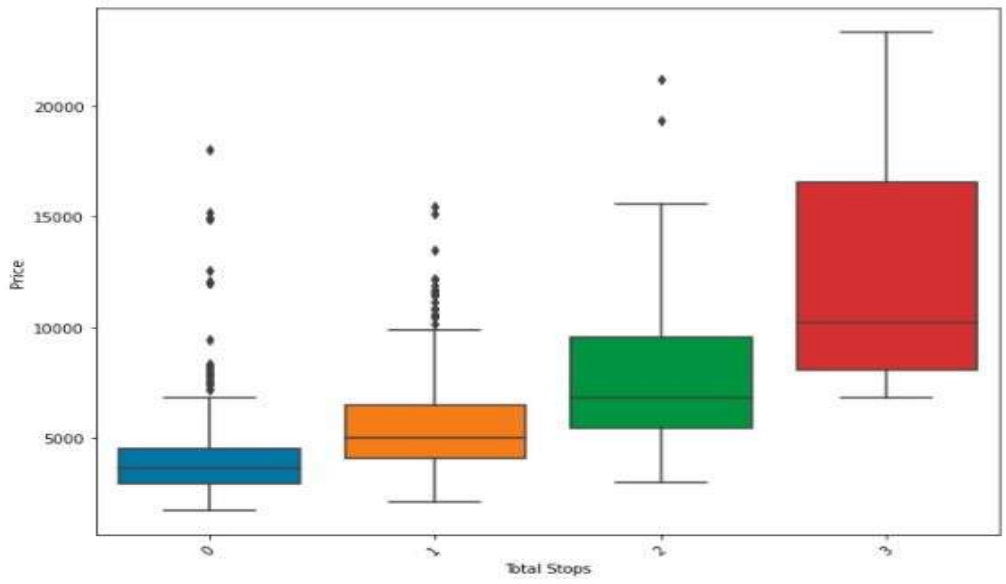
Gradient boosting model is also an ensemble algorithm used for both classification and regression problem. The R2 score of this model is 67.78%.

- Visualizations

The boxplot shows the price range is different in every source and destination. The price is increases with increase in the number of total stops.

```
plt.figure(figsize=(10,8))
plt.xticks(rotation='45')
sns.boxplot(x='Total Stops',y='Price',data=df.sort_values('Price',ascending=False))
```

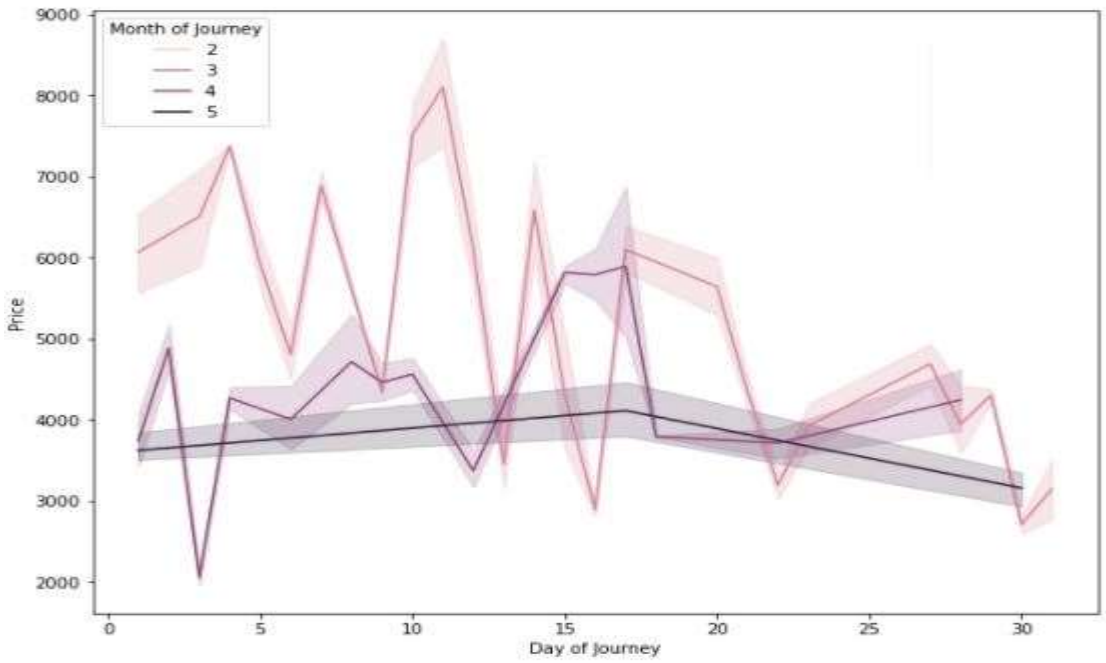
```
<AxesSubplot:xlabel='Total Stops', ylabel='Price'>
```



From the line plot between Day of journey and price, the price for ticket in 3rd month is higher compared to 4th and 5th month. The price is high when the departure date is near. The price is also varies over time.

```
plt.figure(figsize=(10,8))
sns.lineplot(x='Day of Journey',y='Price',hue='Month of Journey',data=df)
```

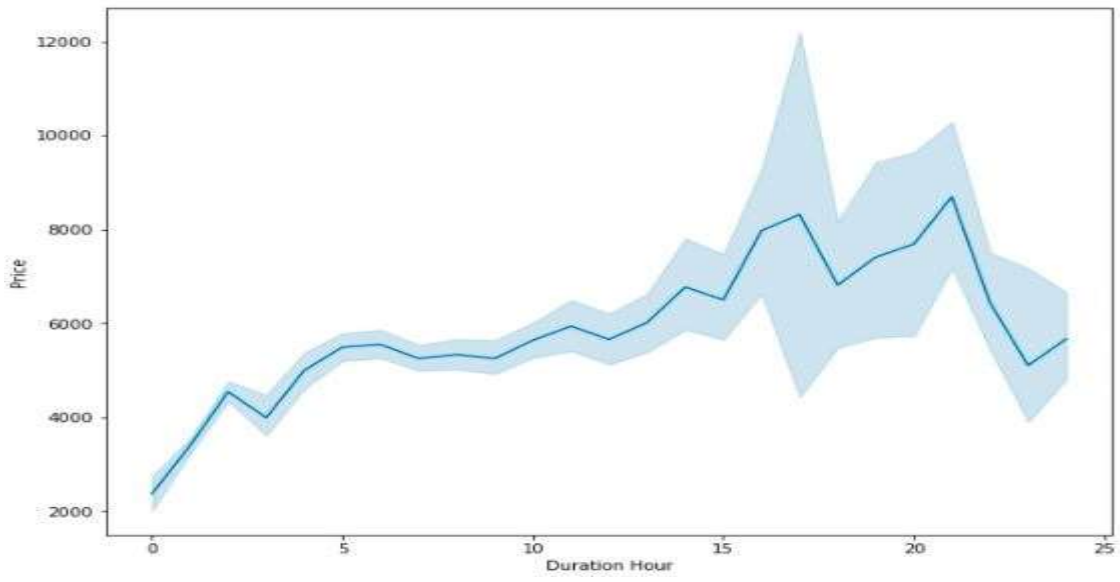
```
<AxesSubplot:xlabel='Day of Journey', ylabel='Price'>
```



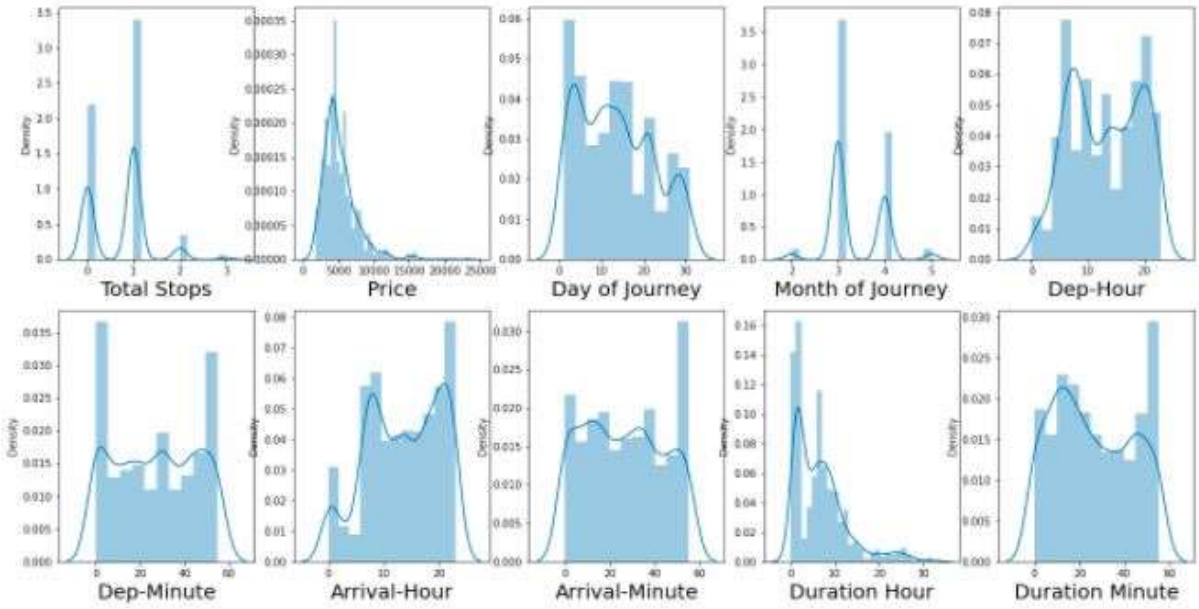
The line plot between Duration Hour and price shows that as the Duration hour increases price is also increasing.

```
plt.figure(figsize=(10,8))
sns.lineplot(x='Duration Hour',y='Price',data=df)
```

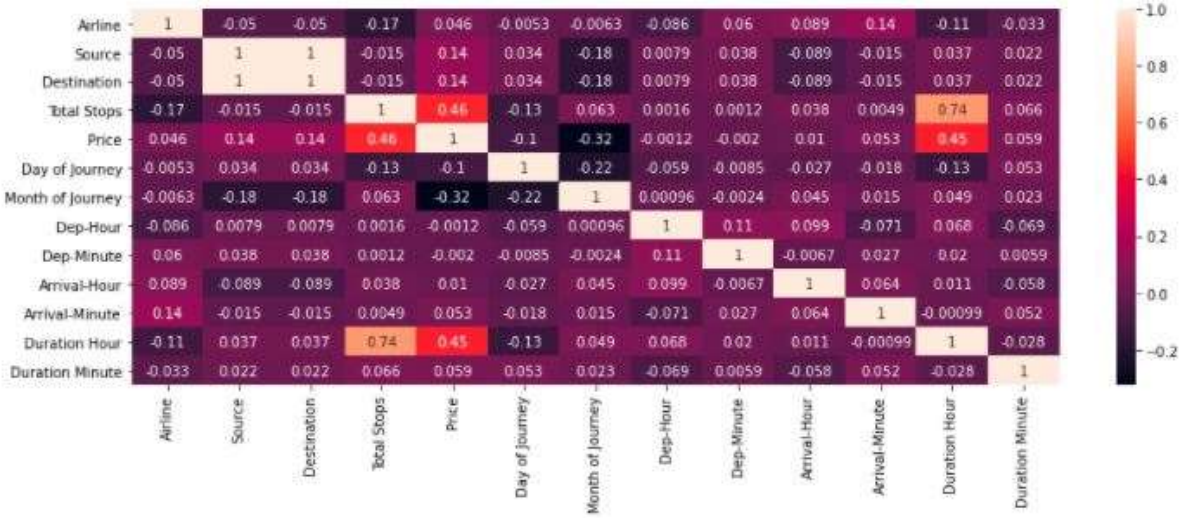
```
<AxesSubplot:xlabel='Duration Hour', ylabel='Price'>
```



The data is normally distributed and skewness is not present in the dataset.



From the correlation plotting the Price is positively correlated to Airline, Source, Destination, Total stops, Arrival Hour, Arrival-minute, Duration Hour and Duration minute. The Price negatively correlated to Day of Journey, Month of Journey, Dep-Hour and Dep-Minute.



- Interpretation of the Results

The price of the flight ticket is mainly depending on the Duration Hour, Total stops, Airline and Source. These features will help to find out the price of the flight ticket. The remaining features plays less role in predicting the price.

CONCLUSION

- Key Findings and Conclusions of the Study

  The price range of the fight is different for every airline, the price also depending on the number of total stops and duration hour as the stops and duration hour increases price also increases. The price increase when the departure date is near. The customer can save more price by taking flights with less stops, less duration hour and far departure days.

- Learning Outcomes of the Study in respect of Data Science

  For predicting the price of the flight ticket different machine learning algorithms are used, after hyper parameter tuning decision tree and gradient boosting model shows better result. From these 2 model decision tree regressor model gives the best result with R2 score 71.62%. The price of a flight ticket mainly depending on total stops, duration hour and departure day.