FLIP ROBO

Flight Price Prediction

Submitted by:

Aditya Kissen Mitra

ACKNOWLEDGMENT

Websites I took help from are-

https://www.analyticsvidhya.com/blog/2022/01/flight-fare-prediction-using-machine-learning/

https://www.yatra.com

INTRODUCTION

- Business Problem Framing

  We need to find out the fare of the flight from one location to another at different conditions and dates

- Conceptual Background of the Domain Problem

  The price of fare is varying with times and the difference between the booking date and travel date.

- Review of Literature

  To find out how the fare of the flight is increasing/decreasing I used different plotting techniques and done my analysis based on the graphs.

- Motivation for the Problem Undertaken

  The motivation behind this problem undertaken is that to find out the flight fare and when it is increasing/decreasing and to guide the user to book the flight when the fare is less.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

When I extracted the data from yatra.com I figured out that most of the fields are of object type so I changed the object type to integer type for different columns.

I plot boxplot between the price of the flight and airline and can conclude that Air India has the most outliers in terms of price.

I plot cat plot between the price of the flight and the destination to which the passenger is travelling and figured out that Mumbai has the most outliers and Goa has the least.

I plot count plot for journey in a month vs several flights and got to see that June has the most number of flights

I plot graph between the type of airline and count of flights can see that Indigo has the most flight boarded

I plot the box plot with the help of a cat plot between the price of the flight and the day to which the passenger is travelling and figured out that Wednesday has the most outliers and Friday has the least

I Plot the box plot with the help of a cat plot between the price of the flight and number of stops to reach the destination and figured out that price is increasing with number of stops.

I plot the scatter plot between the price of the flight and Destination and figured out that flight to Mumbai is most costly.

- Data Sources and their formats

I used different plotting techniques to perform my analysis. It consists of scatter plot, bar graph, count plot, cat plot and box plot.

- ## Data Preprocessing Done

I checked the Null values but found out that no null values present in the dataset.

The departure time, Arrival time was in object type so I convert them to integer type using to_datetime.

For total flight time it was in format hours and minutes together so I convert that columns to total minutes and create a new columns with it and dropped the old column

Month column was in object type with short text example Jan, Feb so I convert it to numbers

Same as Months I convert the days columns to integer as well

- ## Data Inputs- Logic- Output Relationships

The columns was in object type so I convert the columns to integer type and done my analysis.

- ## Hardware and Software Requirements and Tools Used

I used jupyter notebook for my extracting of data, analysis of data and done the machine learning as well.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  After doing analysis I used label encoder to convert the remaining object type columns to integer

  I controlled the skewness using log1() function

  I checked VIF and it was under control (+10/-10)

  I checked the heat-map as well for correlation and found out that no columns are highly correlated to each other.

  Then I perform the scaling of data using standard scalar

- Testing of Identified Approaches (Algorithms)

  Listing down all the algorithms used for the training and testing.

  Algorithms used-

  Linear Regression

  Ridge

  Gradient Boosting

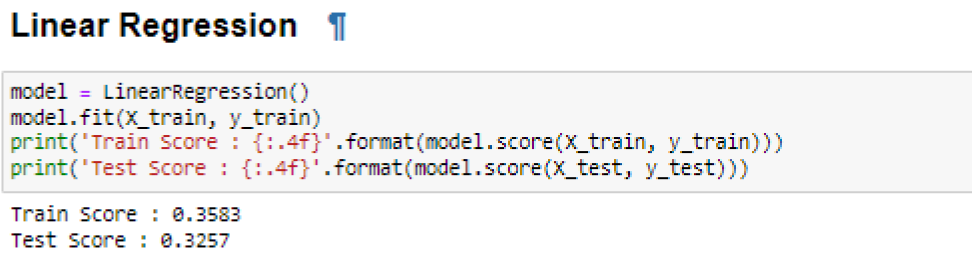
  K-Nearest Neighbours

  Random Tree

  Decision Tree

- Run and Evaluate selected models

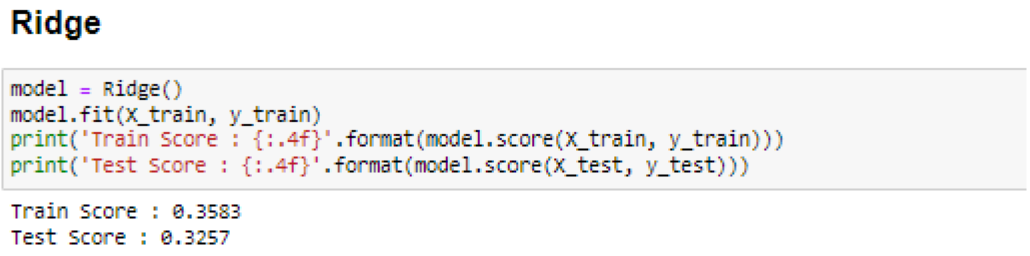
  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.



Linear Regression ¶

```
model = LinearRegression()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

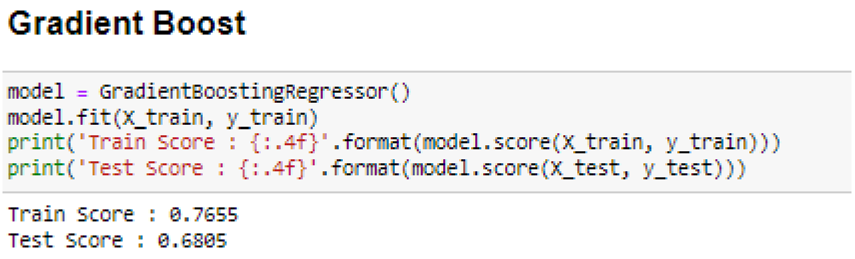
```
Train Score : 0.3583
Test Score : 0.3257
```



Ridge

```
model = Ridge()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

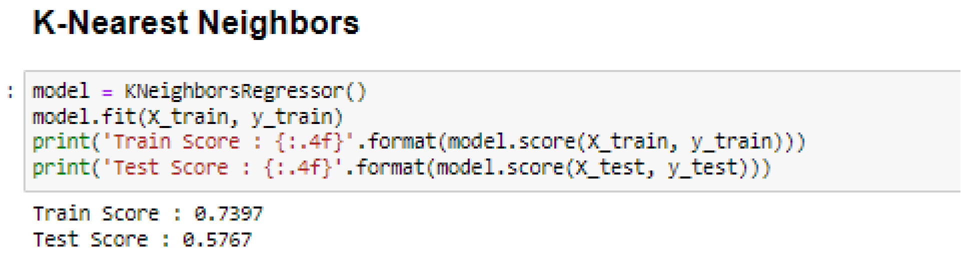
```
Train Score : 0.3583
Test Score : 0.3257
```



Gradient Boost

```
model = GradientBoostingRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

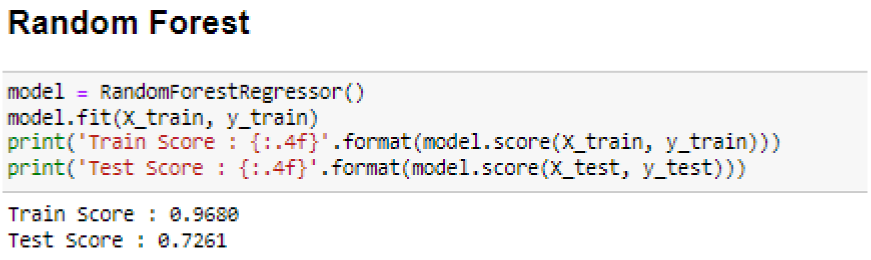
```
Train Score : 0.7655
Test Score : 0.6805
```



K-Nearest Neighbors

```
model = KNeighborsRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

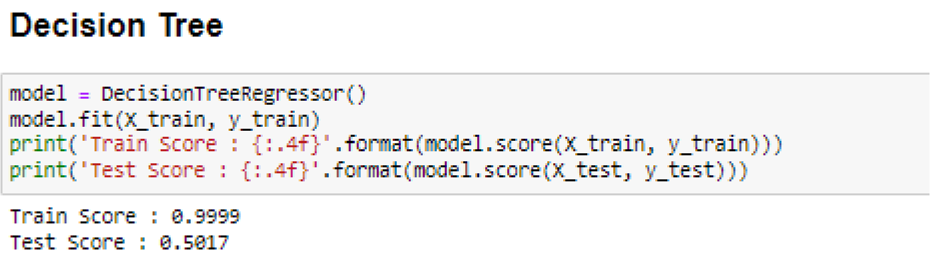
```
Train Score : 0.7397
Test Score : 0.5767
```



Random Forest

```
model = RandomForestRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.9680
Test Score : 0.7261
```



Decision Tree

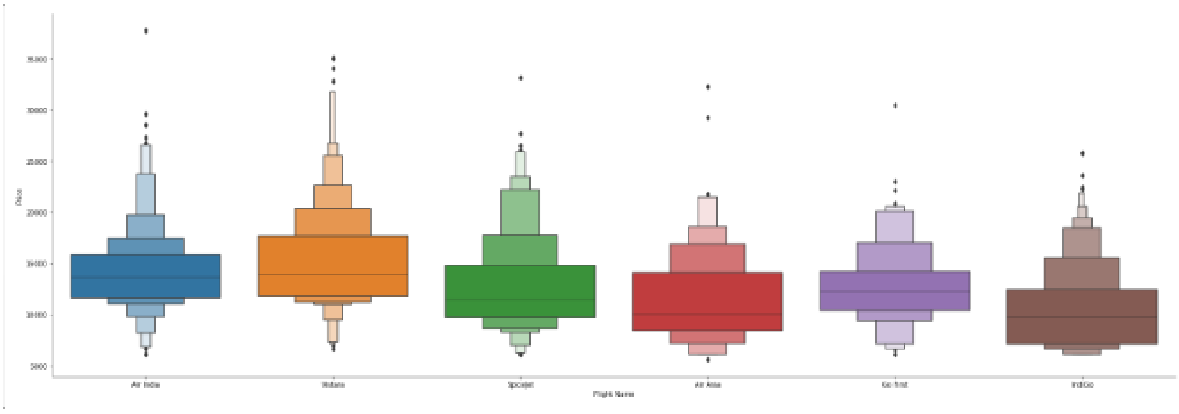
```
model = DecisionTreeRegressor()
model.fit(X_train, y_train)
print('Train Score : {:.4f}'.format(model.score(X_train, y_train)))
print('Test Score : {:.4f}'.format(model.score(X_test, y_test)))
```

```
Train Score : 0.9999
Test Score : 0.5017
```

- Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

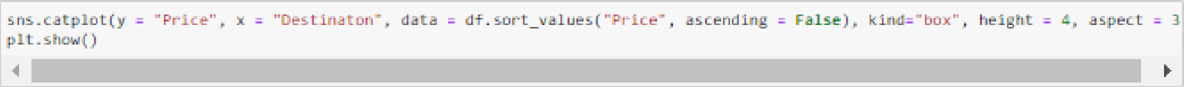
If different platforms were used, mention that as well.



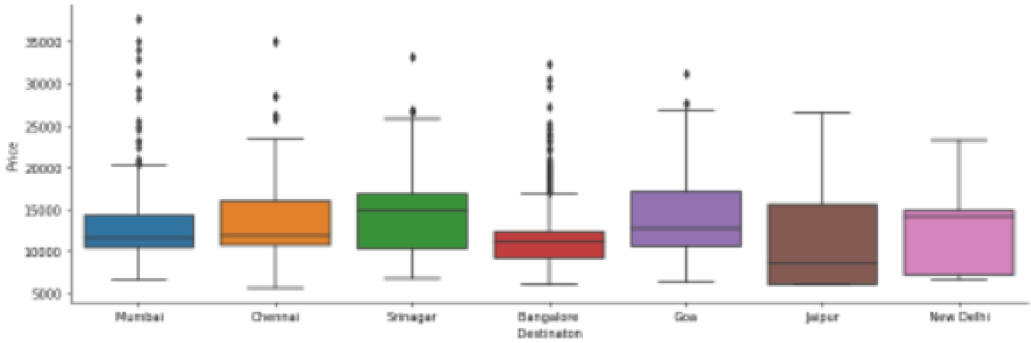
Observation-

Here with the help of the cat plot we are trying to plot the boxplot between the price of the flight and airline and we can conclude that Air India has the most outliers in terms of price.

Plotting Box plot for Price vs Destination



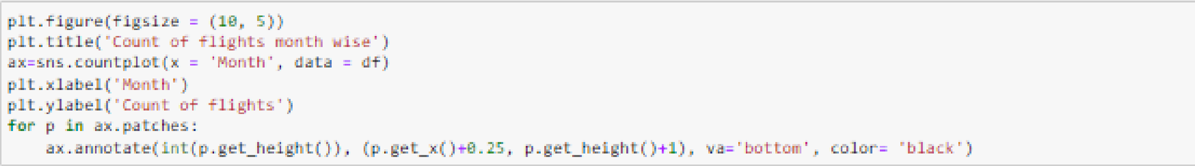
```
sns.catplot(y = "Price", x = "Destinaton", data = df.sort_values("Price", ascending = False), kind="box", height = 4, aspect = 3
plt.show()
```



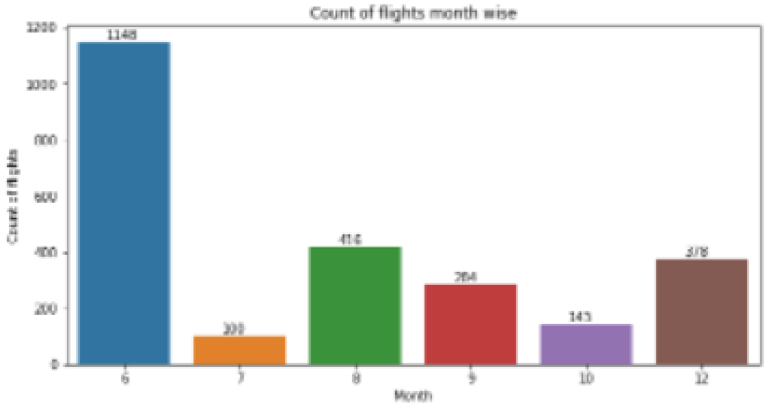
Observation-

Here we are plotting the box plot with the help of a cat plot between the price of the flight and the destination to which the passenger is travelling and figured out that Mumbai has the most outliers and Goa has the least.

Plotting Bar chart for Months (Duration) vs Number of Flights



```
plt.figure(figsize = (10, 5))
plt.title('Count of flights month wise')
ax=sns.countplot(x = 'Month', data = df)
plt.xlabel('Month')
plt.ylabel('Count of flights')
for p in ax.patches:
    ax.annotate(int(p.get_height()), (p.get_x()+0.25, p.get_height()+1), va='bottom', color= 'black')
```

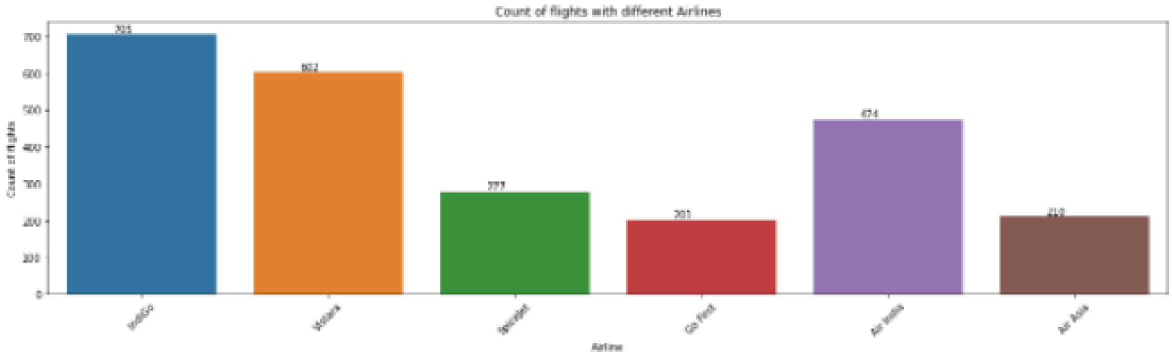


Observation-

Here in the above graph we have plotted the count plot for journey in a month vs several flights and got to see that June has the most number of flights.

Plotting Bar chart for Types of Airline vs Number of Flights

```
plt.figure(figsize = (20,5))
plt.title('Count of flights with different Airlines')
ax=sns.countplot(x = 'Flight Name', data =df)
plt.xlabel('Airline')
plt.ylabel('Count of flights')
plt.xticks(rotation = 45)
for p in ax.patches:
    ax.annotate(int(p.get_height()), (p.get_x()+0.25, p.get_height()+1), va='bottom', color= 'black')
```

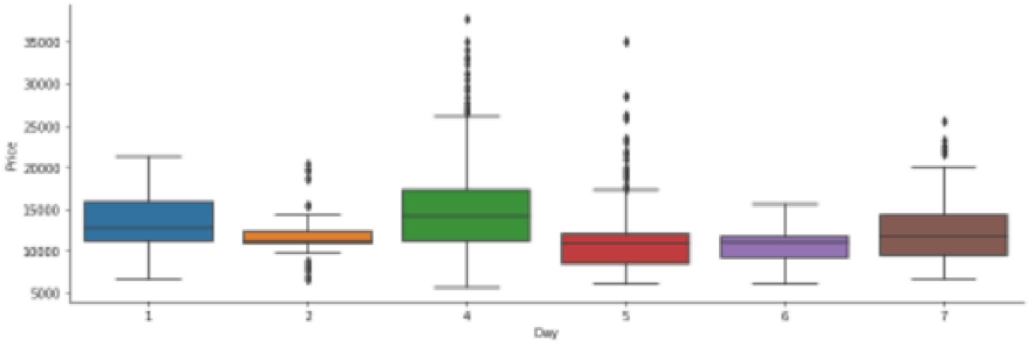


Observation-

Now from the above graph we can see that between the type of airline and count of flights we can see that Indigo has the most flight boarded

Plotting Box plot for Price vs Day

```
sns.catplot(y = "Price", x = "Day", data = df.sort_values("Price", ascending = False), kind="box", height = 4, aspect = 3)
plt.show()
```

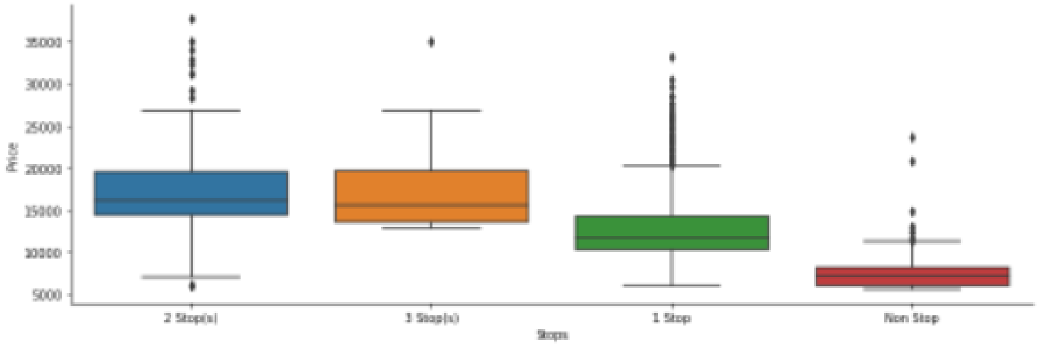


Observation-

Here we are plotting the box plot with the help of a cat plot between the price of the flight and the day to which the passenger is travelling and figured out that Wednesday has the most outliers and Friday has the least.

Plotting Box plot for Price vs Stops

```
sns.catplot(y = "Price", x = "Stops", data = df.sort_values("Price", ascending = False), kind="box", height = 4, aspect = 3)
plt.show()
```



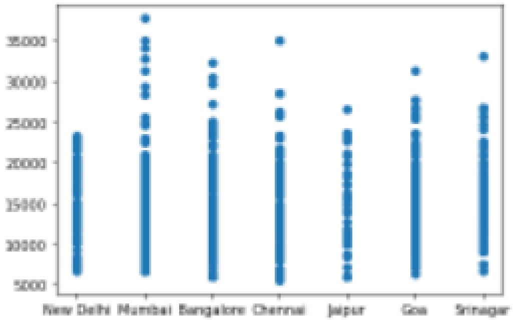
Observation-

Here we are plotting the box plot with the help of a cat plot between the price of the flight and number of stops to reach the destination and figured out that price is increasing with number of stops.

Plotting Ticket Prices VS Destination ¶

```
[ ]: x = df['Destinaton']
     y = df['Price']

     plt.scatter(x, y)
     plt.show()
```



Observation-

Here we are plotting the scatter plot between the price of the flight and Destination and figured out that flight to Mumbai is most costly.

CONCLUSION

- Key Findings and Conclusions of the Study

The departure time, Arrival time was in object type so I convert them to integer type using to_datetime.

For total flight time it was in format hours and minutes together so I convert that columns to total minutes and create a new columns with it and dropped the old column

Month column was in object type with short text example Jan, Feb so I convert it to numbers

Same as Months I convert the days columns to integer as well

- Limitations of this work and Scope for Future Work

For further studies I can work on removing more outliers and try to increase the accuracy more by hyper tuning more.