PROJECT NAME:-

[FLIGHT PRICE PREDICTION]

SUBMITTED BY:

SHIVAM SHARMA

ACKNOWLEDGMENT

- **Yatra.com**
- **Makemytrip.com**

INTRODUCTION

Problem Statement:

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)

So, we have to work on a project where we collect data of flight fares with other features and work to make a model to predict fares of flights.

Data Collection Phase:-

We have to scrape at least 1500 rows of data. We can scrape more data as well, it’s up to us, More the data better the model In this section we have to scrape the data of flights from different websites (yatra.com, skyscanner.com, official websites of airlines, etc). The number of columns for data doesn’t have limit, it’s up to us and our creativity. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. We can make changes to it, we can add or we can remove some columns, it completely depends on the website from which we are fetching the data..

Model Building Phase:-

After collecting the data, we need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Review of Literature

THE TOPIC IS ABOUT THE PREDICTING THE FLIGHT PRICE WHICH IS FETCHED FROM THE MAKEMYTRIP.COM AS WE CAN MAKE A MODEL SO THAT THEY CAN EASILY PREDICT THE PRICE OF THE FLIGHT AS THEY CAN TAKE ADVANTAGE AND CAN SAVE MONEY BY BOOKING THE FLIGHT ON THE DATES WHICH HAVING LOW PRICE..

Motivation for the Problem Undertaken

So, we have to work on a project where we collect data of flight fares with other features and work to make a model to predict fares of flights.

Do airfares change frequently?

Do they move in small increments or in large jumps?

Do they tend to go up or down over time?

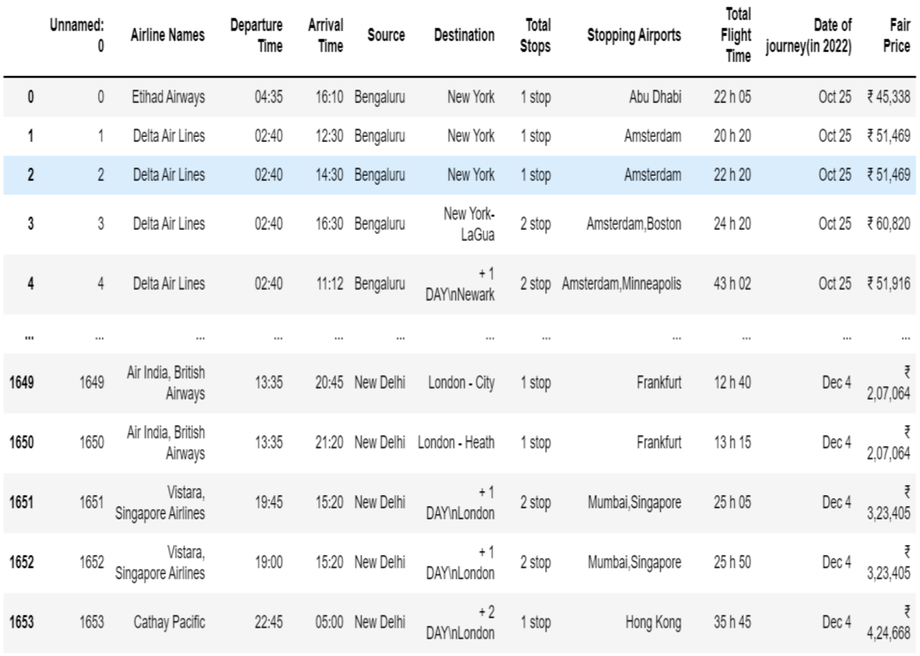
What is the best time to buy so that the consumer can save the most by taking the least risk?

Does price increase as we get near to departure date?
Is Indigo cheaper than Jet Airways?

Are morning flights expensive?

Data Sources and their format

- YATRA.COM
- MAKEMYTRIP.COM



	Unnamed: 0	Airline Names	Departure Time	Arrival Time	Source	Destination	Total Stops	Stopping Airports	Total Flight Time	Date of journey(in 2022)	Fair Price
0	0	Etihad Airways	04:35	16:10	Bengaluru	New York	1 stop	Abu Dhabi	22 h 05	Oct 25	₹ 45,338
1	1	Delta Air Lines	02:40	12:30	Bengaluru	New York	1 stop	Amsterdam	20 h 20	Oct 25	₹ 51,469
2	2	Delta Air Lines	02:40	14:30	Bengaluru	New York	1 stop	Amsterdam	22 h 20	Oct 25	₹ 51,469
3	3	Delta Air Lines	02:40	16:30	Bengaluru	New York-LaGua	2 stop	Amsterdam,Boston	24 h 20	Oct 25	₹ 60,820
4	4	Delta Air Lines	02:40	11:12	Bengaluru	+ 1 DAY\nNewark	2 stop	Amsterdam,Minneapolis	43 h 02	Oct 25	₹ 51,916
...	...	...	...	...	...	...	...	...	...	...	...
1649	1649	Air India, British Airways	13:35	20:45	New Delhi	London - City	1 stop	Frankfurt	12 h 40	Dec 4	₹ 2,07,064
1650	1650	Air India, British Airways	13:35	21:20	New Delhi	London - Heath	1 stop	Frankfurt	13 h 15	Dec 4	₹ 2,07,064
1651	1651	Vistara, Singapore Airlines	19:45	15:20	New Delhi	+ 1 DAY\nLondon	2 stop	Mumbai,Singapore	25 h 05	Dec 4	₹ 3,23,405
1652	1652	Vistara, Singapore Airlines	19:00	15:20	New Delhi	+ 1 DAY\nLondon	2 stop	Mumbai,Singapore	25 h 50	Dec 4	₹ 3,23,405
1653	1653	Cathay Pacific	22:45	05:00	New Delhi	+ 2 DAY\nLondon	1 stop	Hong Kong	35 h 45	Dec 4	₹ 4,24,668

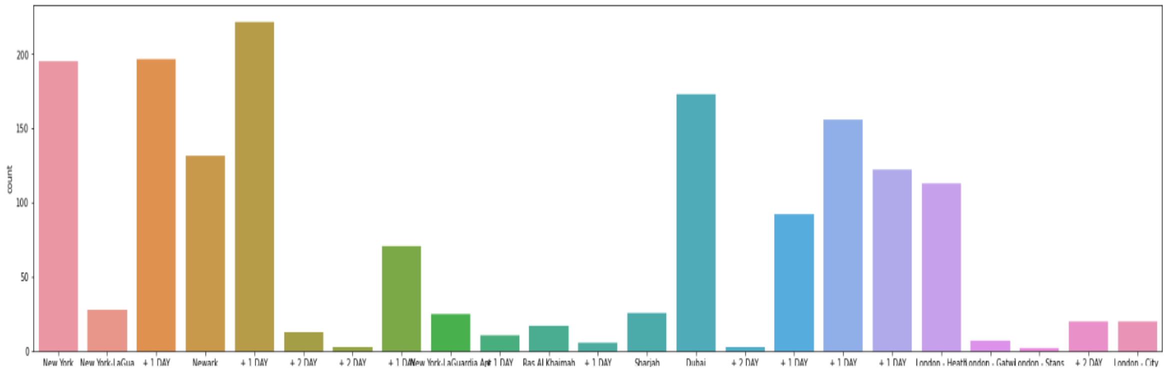
Data Preprocessing and EDA

The steps followed for the cleaning and EDA of the data:-

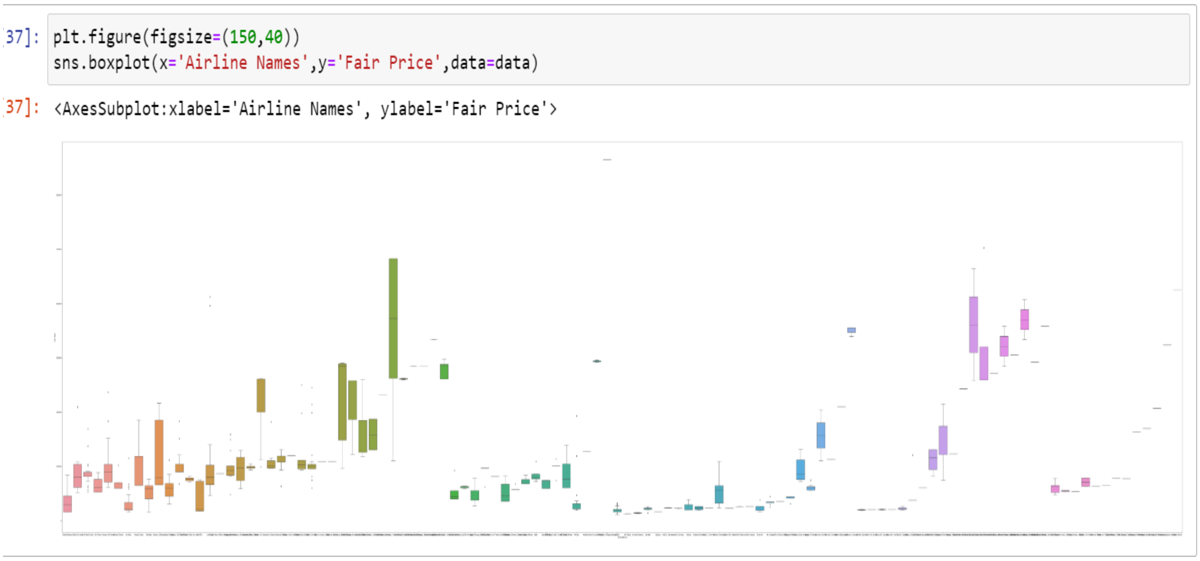
1) First we analyse the data by simply DATA.INFO() method..

2) Then we manipulate the Dates and make it for ML use so that it can be understandable my ML.
3) After dates we drop some unneccesary columns like(Date of jouney as we have dates and months separately in diff colummns and dropped unnamed:0 as this column only provide indexing)
4) Then after we replaced some strings from the columns so that we can make them int data type.
5) After that we analyse the data with countplot visualization techniques with column on x axis by which it is easy to interpret the data.

```
Name: Destination, dtype: int64
```



6) Then we replace some columns value which has object datatype so that they can be easily understandable by machine i.e- int datatypes..
7) After that we plot a graph of Boxplot , with the help of boxplot we analyse the Price range for many Airlines..



```
37]: plt.figure(figsize=(150,40))
     sns.boxplot(x='Airline Names',y='Fair Price',data=data)

37]: <AxesSubplot:xlabel='Airline Names', ylabel='Fair Price'>
```

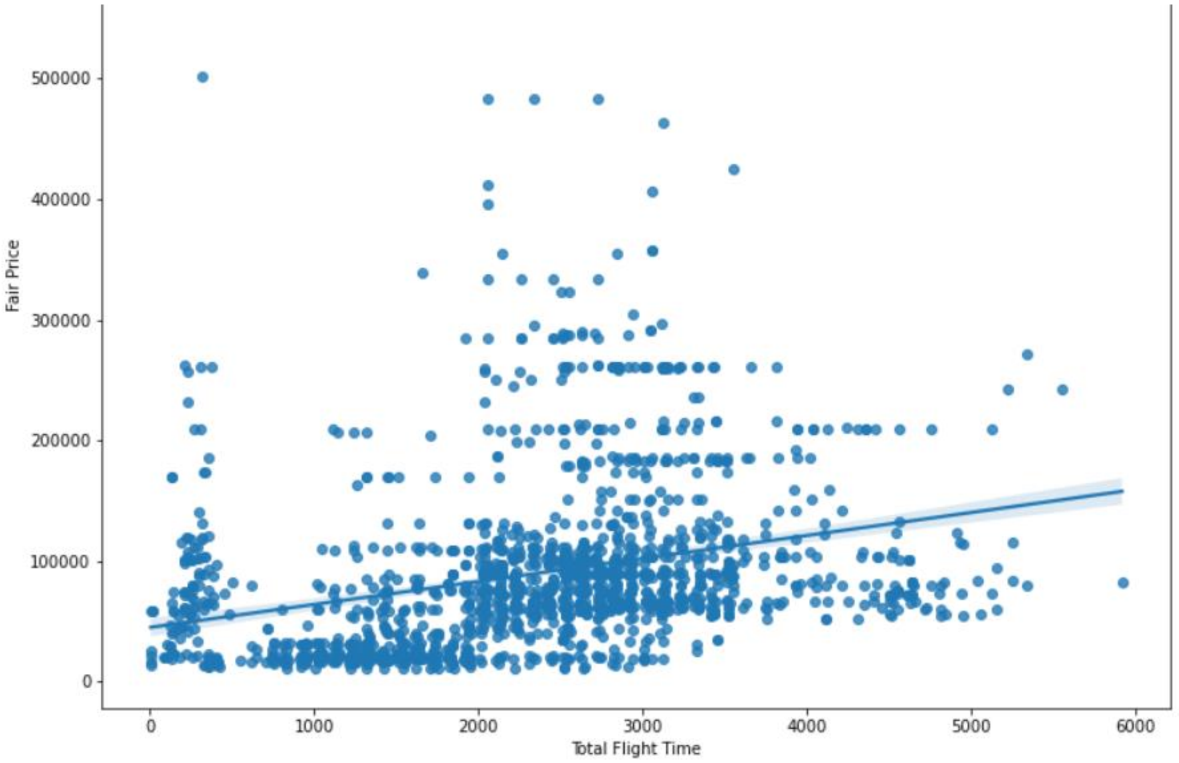
This graph tells about the price range of diff. airlines

- LUFTHANSA AND UNITED AIRLINE SHOWS THE BIGGEST PRICE RANGE FOR THE FLIGHTS

- INDIGO , SPICEJET , AIR INDIA , QATAR AIRWAYS(ETC.) shows the lowest price range

8) After that we plot a graph of Regplot , with the help of Regplot we analyse the Price trend with TOTAL TIME TAKEN FOR THE FLIGHT AND DATES/MONTHS for many Airlines..So that it can be easy for the customers to they tend to go up or down over time?

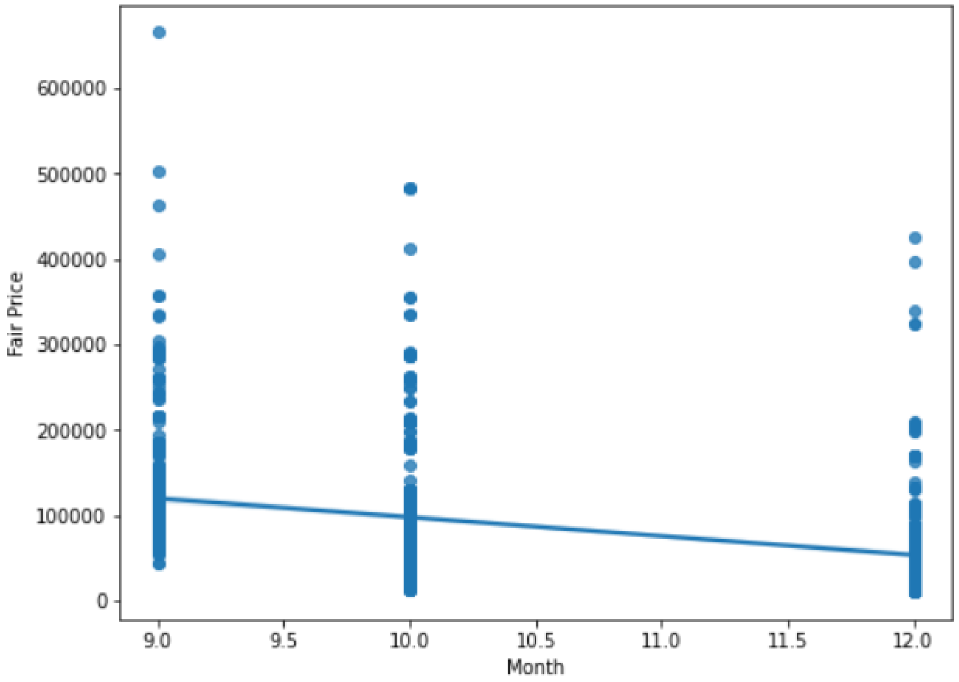
- **First with Total time taken**



It clarifies from the graph as the Total time Flight increases then the price of the Flight will increases.

- **<u>With DATES/MONTHS</u>**

```
: <AxesSubplot:xlabel='Month', ylabel='Fair Price'>
```



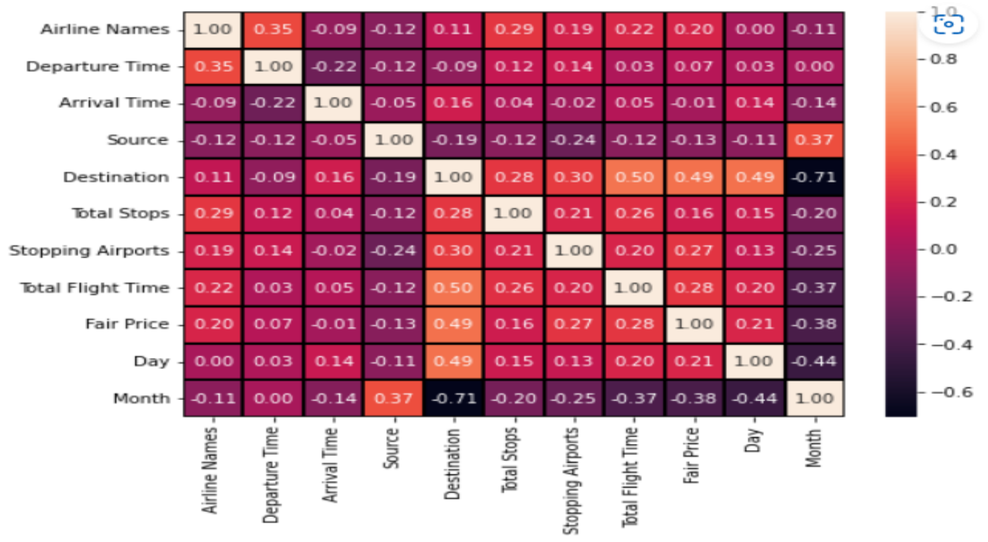
FROM THIS GRAPH WE CAN UNDERSTAND THE PRICE OF THE TREND

AS WE BOOKED A TICKET NEARBY OUR CURRENT DATE THEN THE PRICE IS VERY HIGH AS COMPARE TO THE DATES OF COMING MONTHS ..

IT MEANS PRICE WILL BE CHEAPER IF WE BOOKED A TICKET BEFORE 2 3 MONTHS..

9) Then after we use encoding techniques for wncoding some object datatype anad make int as datatype now our data is good for ML .
10) Now we will check the correlation between the independent variables and between the independent variables and dependent variables.
11) Then we Plot a Heatmap to check clearly the correlations between all the variables.

```
<AxesSubplot:>
```



	Airline Names	Departure Time	Arrival Time	Source	Destination	Total Stops	Stopping Airports	Total Flight Time	Fair Price	Day	Month
Airline Names	1.00	0.35	-0.09	-0.12	0.11	0.29	0.19	0.22	0.20	0.00	-0.11
Departure Time	0.35	1.00	-0.22	-0.12	-0.09	0.12	0.14	0.03	0.07	0.03	0.00
Arrival Time	-0.09	-0.22	1.00	-0.05	0.16	0.04	-0.02	0.05	-0.01	0.14	-0.14
Source	-0.12	-0.12	-0.05	1.00	-0.19	-0.12	-0.24	-0.12	-0.13	-0.11	0.37
Destination	0.11	-0.09	0.16	-0.19	1.00	0.28	0.30	0.50	0.49	0.49	-0.71
Total Stops	0.29	0.12	0.04	-0.12	0.28	1.00	0.21	0.26	0.16	0.15	-0.20
Stopping Airports	0.19	0.14	-0.02	-0.24	0.30	0.21	1.00	0.20	0.27	0.13	-0.25
Total Flight Time	0.22	0.03	0.05	-0.12	0.50	0.26	0.20	1.00	0.28	0.20	-0.37
Fair Price	0.20	0.07	-0.01	-0.13	0.49	0.16	0.27	0.28	1.00	0.21	-0.38
Day	0.00	0.03	0.14	-0.11	0.49	0.15	0.13	0.20	0.21	1.00	-0.44
Month	-0.11	0.00	-0.14	0.37	-0.71	-0.20	-0.25	-0.37	-0.38	-0.44	1.00

12) Since the data Is continuous data so we are not checking the skewness and outliers.
13) After these methods we seprate the labels and features in different dataframe so that we can easily apply some methods on the labels and features differently..
14) After this we applied scaling on the features (x) by the help of   Standard Scaler…
15) Now we are ready for the model building……

Algorithm used in this project:-

For building machine learning models there are several models present inside the Sklearn module.

Sklearn provides two types of models i.e. regression and classification. Our dataset's target variable is to predict the sale price of the car. So for this kind of problem we use regression models.

But before the model fitting we have to seprate the predictor and target variable, then we pass this variable to the train_test_split method to create the training set and testing set for the model training and prediction.

We can build as many models as we want to compare the accuracy given by these models and to select the best model among them.

I have selected 5 models:

1. Linear Regression
2. Lasso
3. Random forest Regressor
4. Adaboost Regressor
5. Gradient boosting Regressor

BEST Algo. From these And why?

Gradient boosting Regressor is the best algo from all of these algo which is used in this data to predict because the difference between the cross val score and the accuracy score is minimum for this algo and it also give better **ACCURACY(approx. 80%)** after Hypertuning with GRIDSEARCHCV that's why we USED THIS **Algo.**

- **Save the model for later predictions by the help of pickle..**

# Now my model is ready to predict

CONCLUSIONS

THE DATAFRAME CONTAINS THE DATA WHICH RELATED TO THE FLIGHT PRICE PREDICTION ..

We got our best model i.e **GRADIENT BOOSTING R EGRESSOR with the accuracy score of 80**%.
HERE our model predicts ROOT MEAN SQUARED ER ROR OF 34921.88 THAT IS VERY LOW THAN OTHERS ..

KEYFINDING:-

- **AS The Time Flight increases then the price of the Flight will also increases**
- **AS WE BOOKED A TICKET NEARBY OUR CURRENT DATE THEN THE PRICE IS VERY HIGH AS COMPARE TO THE DATES OF COMING MONTHS ..**
- **IT MEANS PRICE WILL BE CHEAPER IF WE BOOKED A TICKET BEFORE 2 3 MONTHS..**