FLIGHT PRICE PREDICTION

Submitted by:-

ABHISHEK SHAHI

ACKNOWLEDGMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Flip Robo Technologies, Bangalore for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I want to thank my **SME KHUSHBOO GARG** for providing the Dataset and helping us to solve the problem and addressing out our Query in right time. I would like to express my gratitude towards my parents & members of Flip Robo for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to our institute DataTrained & others seen unseen hands which have given us direct & indirect help in completion of this project. With help of their brilliant guidance and encouragement, I was able to complete my tasks properly and were up to the mark in all the tasks assigned. During the process, I got a chance to see the stronger side of my technical and non-technical aspects and also strengthen my concepts.

INTRODUCTION

- ## Business Problem Framing

  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)

- ## Conceptual Background of the Domain Problem

  Airline companies use complex algorithms to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing, and various social factors into account to predict flight prices.

  Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That's why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly

- Review of Literature

As per the requirement of client, I have scrapped the data from online sites and based on that data I have did analysis like for based on which feature of my data prices are changing. and checked the relationship of flight price with all the feature like what flight he should choose.

- Motivation for the Problem Undertaken

I have worked on this on the bases of client requirements and followed all the steps till model deployment.

Analytical Problem Framing

After scrapping my data using selenium I have loaded my data into python with the help of pandas.

```
data=pd.read_csv('/content/Web_Scraped_Flight_Data1.csv')
```

```
data
```

	Unnamed: 0	Unnamed: 0.1	Airline	Source	Destination	Dep_Time	Arrival_Time	Duration	Total_Stops	Price
0	0	0	Jet Airways	Banglore	Delhi	18:55	22:00	3h 5m	non-stop	7229
1	1	1	Multiple carriers	Delhi	Cochin	10:20	01:30 22 May	15h 10m	1 stop	7485
2	2	2	IndiGo	Banglore	Delhi	18:55	21:50	2h 55m	non-stop	4823
3	3	3	Air India	Delhi	Cochin	05:55	07:40 07 Mar	25h 45m	2 stops	14641
4	4	4	SpiceJet	Kolkata	Banglore	06:55	09:30	2h 35m	non-stop	3841
...	...	...	File display	...	...	...	...	...	...	...
1669	1669	1669	IndiGo	Banglore	Delhi	04:00	06:50	2h 50m	non-stop	4423
1670	1670	1670	Jet Airways	Kolkata	Banglore	08:25	18:15	9h 50m	1 stop	10844
1671	1671	1671	Jet Airways	Delhi	Cochin	19:30	12:35 28 Jun	17h 5m	2 stops	13764
1672	1672	1672	Air India	Delhi	Cochin	23:00	19:15 10 Mar	20h 15m	1 stop	11260
1673	1673	1673	Jet Airways	Kolkata	Banglore	16:30	20:45 13 May	28h 15m	1 stop	10844

1674 rows x 10 columns

The size of the data is 1674*8

Data Pre-processing

Checking null values

```
data.isna().sum()
```

```
Unnamed: 0      0
Unnamed: 0.1    0
Airline         0
Source          0
Destination     0
Dep_Time        0
Arrival_Time    0
Duration        0
Total_Stops     0
Price           0
dtype: int64
```

There are no missing values

as there is no null values so I can move forward

- Data Sources and their formats

  I have collected data from web scrapping and I have converted it into csv format

- Data Preprocessing Done

Doing pre-processing where I am dropping some columns and filling missing values in total stops

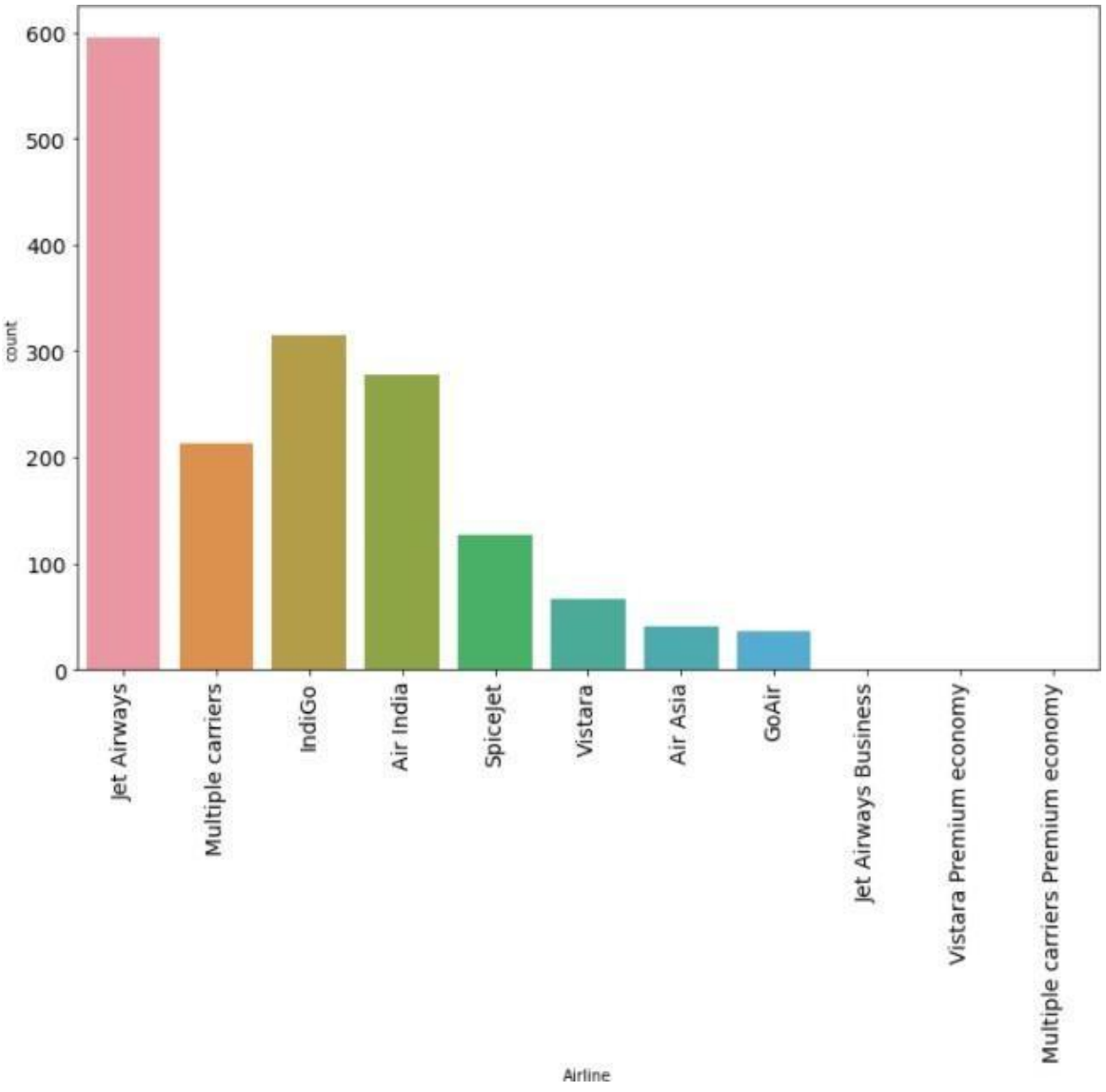
```
def preprocess1(df):
  df['Total_Stops']=df['Total_Stops'].fillna(df['Total_Stops'].mode()[0])
  df=df.drop(['Duration'],axis=1)
  return df
```

```
def preprocess2(df):
  df['Dep_hour']=pd.to_datetime(df['Dep_Time']).dt.hour
  df['Dep_minute']=pd.to_datetime(df['Dep_Time']).dt.minute
  df=df.drop(['Dep_Time'],axis=1)
  df['arrival_hour']=pd.to_datetime(df['Arrival_Time']).dt.hour
  df['arrival_minute']=pd.to_datetime(df['Arrival_Time']).dt.minute
  df=df.drop(['Arrival_Time'],axis=1)
  return df
```

Here I am converting time into hour and minute and also dropping some columns that are not useful for my model.

- Data Inputs- Logic- Output Relationships

I have did EDA to understand the feature relationship.



Obseravtion

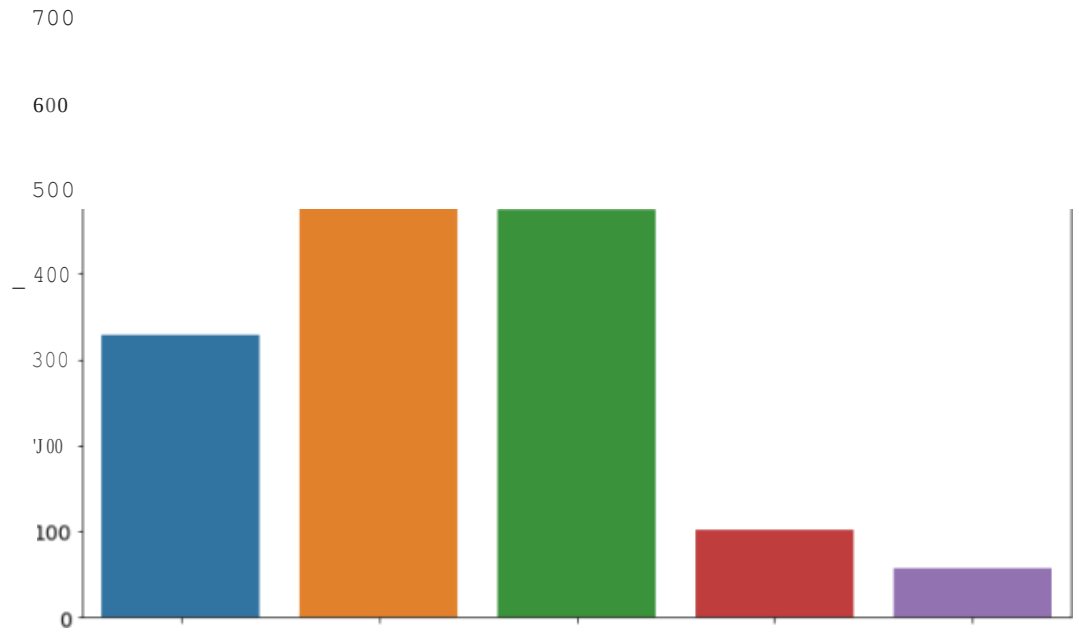
1- Mostly people use to travel with Jet Airways

2- After Jet Airways people use to travel with IndiGo

3- ANd GoAir has the least count why i am saying least count because Vistara,Jet Airways,MCPE,Trujet has no count

Countplot of Source

```
countplot(data['Source
```



Obseravtion

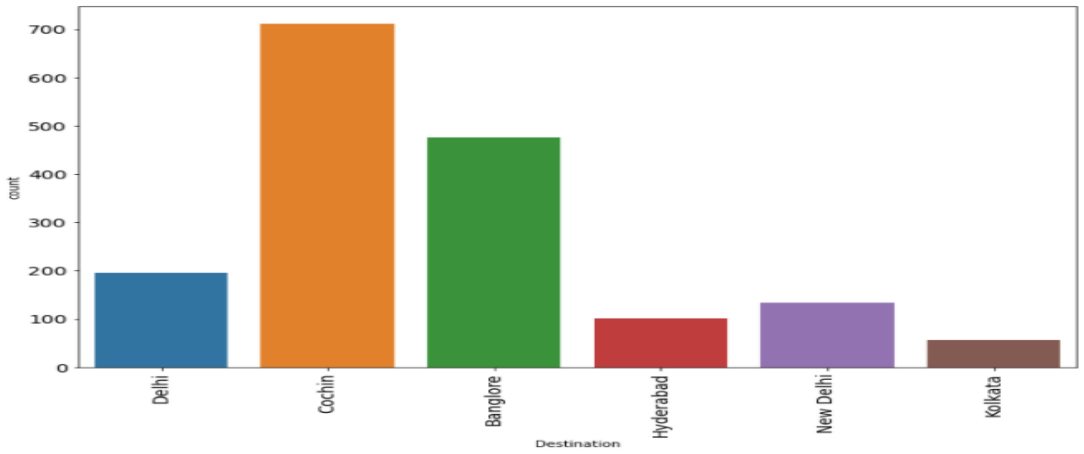
1-Mostly Source has delhi as high count

2-after delhi kolkata has 2nd high count

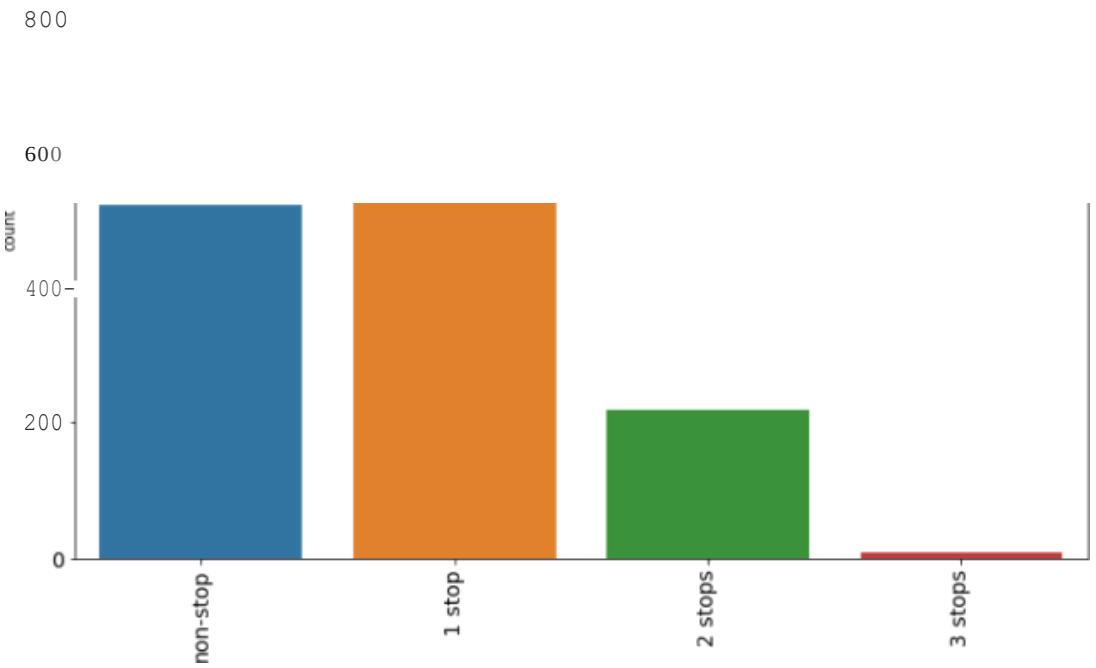
3-and at least chennai means very less people source is chennai

Countplot of destination

```
countplot(data['Destination'])
```



cc min I p1ot: cz I a [' T o I a1 it G : 5 '])



- State the set of assumptions (if any) related to the problem under consideration

  Here, you can describe any presumptions taken by you.

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  I have did Analysis on this data to understand the value of each feature and the contribution of each feature for model creating and effect of all the feature on the prices. And considering all the point I have built a model that can predict the prices.

  ### Testing of Identified Approaches (Algorithms)

  I have trained many model and even evaluate them using all the performance metrics of regression. Here is the screenshot

```
models_result
```

	NAME	Cross_Val_Score	R2_score	Mean_squared_error	Mean_Absolute_Error	RMSE
0	XGB Regressor	0.719294	0.728982	0.068080	0.197141	0.260922
1	ExtraTrees Regressor	0.692417	0.686074	0.078859	0.184097	0.280818
2	RandomForest Regressor	0.714508	0.738310	0.065737	0.178241	0.256392
3	Linear Regression	0.445100	0.422638	0.145034	0.310727	0.380834
4	DecisionTree Regressor	0.637543	0.604147	0.099439	0.206853	0.315340
5	Lasso	-0.006819	-0.003943	0.252192	0.421478	0.502187
6	LIGHT GBM	0.721296	0.730077	0.067805	0.185964	0.260394

  I have make a dataframe of all the model and metrics so here we can see the performance of every model. I have selected LIGHT

GBM as a final model because of its accuracy and performance metrics.

- Key Metrics for success in solving problem under consideration

  I have make a dataframe of all the model and metrics so here we can see the performance of every model. I have selected LIGHT GBM as a final model because of its accuracy and performance metrics.

- Interpretation of the Results

  From the above eda we can easily understand the relationship between features and and we can even see which things are effecting the price of flights.

CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

- Learning Outcomes of the Study in respect of Data Science

  The above research will help our client to study the latest flight price market and with the help of the model built he can easily predict the price ranges of the flight, and also will helps him to understand Based on what factors the fight price is decided.

The limitation of the study is that in the volatile changing market we have taken the data, to be more precise we have taken the data at the time of pandemic and recent data, so when the pandemic ends the market correction might happen slowly. So based on that again the deciding factors of the might change and we have shortlisted and taken these data from the important cities across india, if the customer is from the different city our model might fail to predict the accuracy prize of that flight.