FLIP ROBO

FLIGHT PRICE PRDICTION

Submitted By :

Saurav Kumar

ACKNOWLEDGMENT

I would like to thank Flip Robo Technologies for providing me with the opportunity to work on this project from which I have learned a lot. I am also grateful to Shwetank Mishra for his constant guidance and support .

INTRODUCTION

- Business Problem Framing

As a Business Problem we can see that . if we scrape the data from website like goibibo, yatra.com and etc .and our task is to predict the Price . so it is very good for business because if we know the price of next flight so we can assume the everything ho many tax we have to pay , how many price we have to take from customer . and its also good for customers because customers can know the price of the flight which are in next month.

- Conceptual Background of the Domain Problem

As we can se that in Problem statement we have to predict the price of the flight Using Scrap data from any web site.

So, Here When we scrap the data .and when we plot in the Jupyter Notebook We can see that Maximum Flights Come from Vistara and Indigo airline . and least is Alliance Air .

So using all the Air Line Details we have to build a Regression Model .

- Review of Literature

So , If any one wants to book a flight ticket and He he don’t know which time the ticket is cheaper and which time ticker is costly . The cheapest available ticket on a given flight gets more and less expensive over time. So when we book ticket

for next 2-3 Hours so definitely our flight ticket is costly and we book ticket for next moth we can find that our ticket is cheaper . and we can also find that if there is a festival so ticket price bumps up. And this is a attempt to maximize revenue by a Airline Company. So using all the scrap data we have to build a machine learning model . and Price is our Dependent Variable and left are Independent Variable .

- Motivation for the Problem Undertaken

So, Here after Scraping the data . our motivation is to find the Price of Flight . and before Building a Model we have to know that Why we find the price ? Our motivation is to find the price because it help the airline company to what is the price for next six month of the flight. And if we build the right model it helps us in to how many seat will be booked and how much we have to charge for 1 seat . and its also help to the customer to know the expected price for flights . and more imp thing how much tax pay to the govt in next 6 month ..

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

➢ **So , In Our data set we have 3731 Rows and 11 Columns**

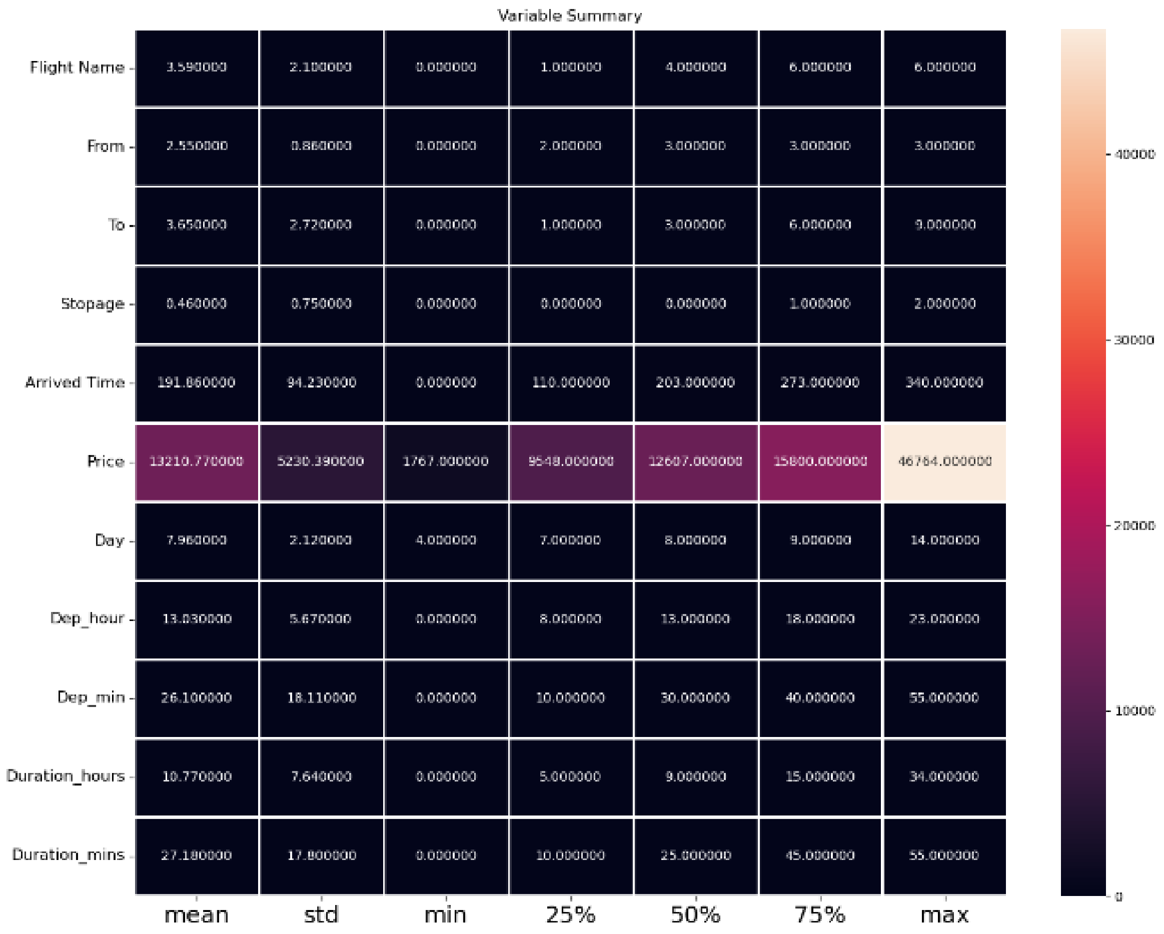
➢ **There is no-null values .**

➢ **All the Columns have Object data type. We have to convert into Date time and Float Datatype.**

- Data Sources and their formats

➢ **So, all the data scrape from yatra.com .**

➢ **All the scraped data save in CSV formet**



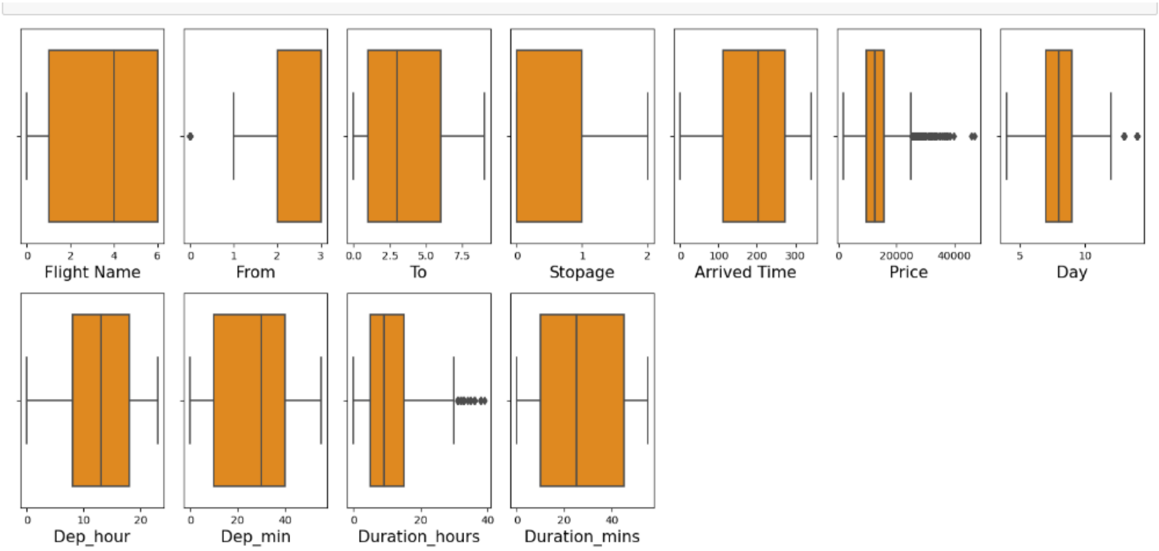
Variable Summary

	mean	std	min	25%	50%	75%	max
Flight Name	3.590000	2.100000	0.000000	1.000000	4.000000	6.000000	6.000000
From	2.550000	0.860000	0.000000	2.000000	3.000000	3.000000	3.000000
To	3.650000	2.720000	0.000000	1.000000	3.000000	6.000000	9.000000
Stopage	0.460000	0.750000	0.000000	0.000000	0.000000	1.000000	2.000000
Arrived Time	191.860000	94.230000	0.000000	110.000000	203.000000	273.000000	340.000000
Price	13210.770000	5230.390000	1767.000000	9548.000000	12607.000000	15800.000000	46764.000000
Day	7.960000	2.120000	4.000000	7.000000	8.000000	9.000000	14.000000
Dep_hour	13.030000	5.670000	0.000000	8.000000	13.000000	18.000000	23.000000
Dep_min	26.100000	18.110000	0.000000	10.000000	30.000000	40.000000	55.000000
Duration_hours	10.770000	7.640000	0.000000	5.000000	9.000000	15.000000	34.000000
Duration_mins	27.180000	17.800000	0.000000	10.000000	25.000000	45.000000	55.000000

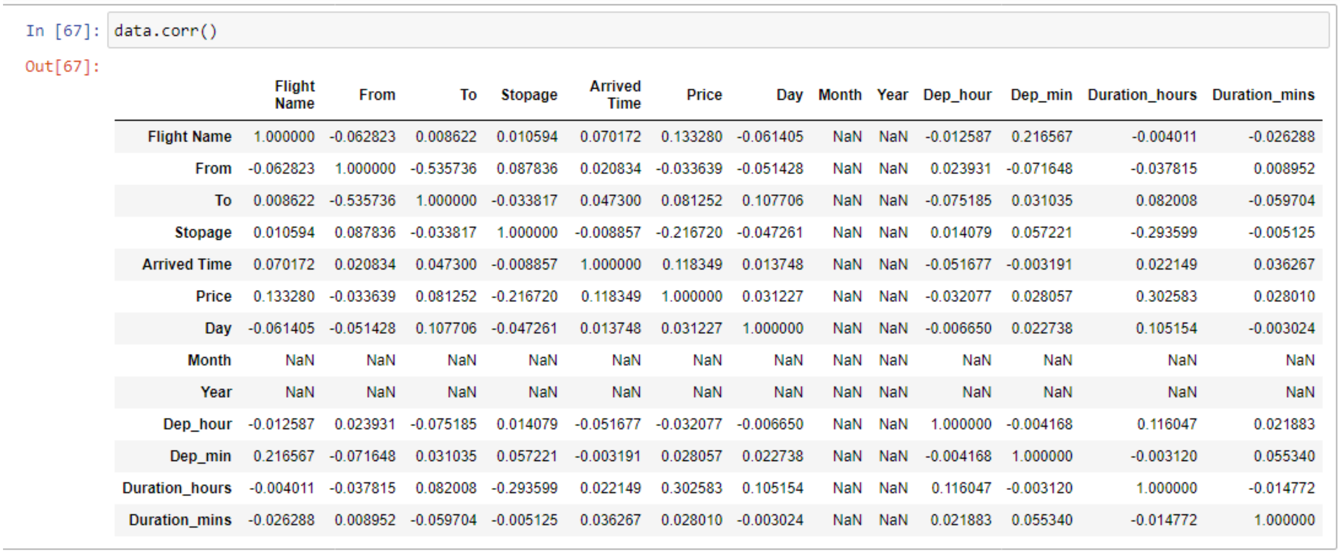
- Data Preprocessing Done

➢ **As we all know our maximum data are in Object data. So our first step is encode all the data using label encoder .**

After Encoding : We can check for outliers.We find some outliers but we don’t consider as a outlier because of domain knowledge of flight.



➢ **Checking the Correlation .**

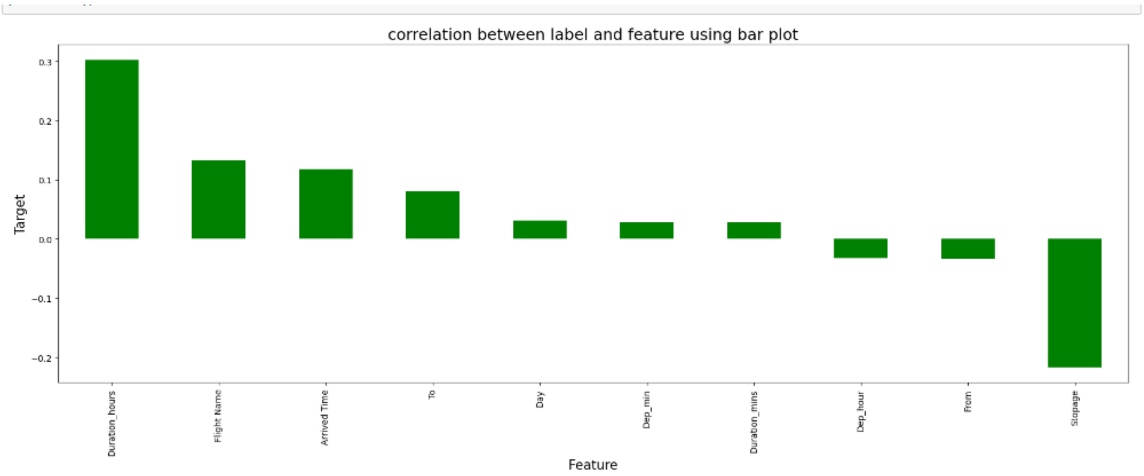


```
In [67]: data.corr()
Out[67]:
```

	Flight Name	From	To	Stopage	Arrived Time	Price	Day	Month	Year	Dep_hour	Dep_min	Duration_hours	Duration_mins
Flight Name	1.000000	-0.062823	0.008622	0.010594	0.070172	0.133280	-0.061405	NaN	NaN	-0.012587	0.216567	-0.004011	-0.026288
From	-0.062823	1.000000	-0.535736	0.087836	0.020834	-0.033639	-0.051428	NaN	NaN	0.023931	-0.071648	-0.037815	0.008952
To	0.008622	-0.535736	1.000000	-0.033817	0.047300	0.081252	0.107706	NaN	NaN	-0.075185	0.031035	0.082008	-0.059704
Stopage	0.010594	0.087836	-0.033817	1.000000	-0.008857	-0.216720	-0.047261	NaN	NaN	0.014079	0.057221	-0.293599	-0.005125
Arrived Time	0.070172	0.020834	0.047300	-0.008857	1.000000	0.118349	0.013748	NaN	NaN	-0.051677	-0.003191	0.022149	0.036267
Price	0.133280	-0.033639	0.081252	-0.216720	0.118349	1.000000	0.031227	NaN	NaN	-0.032077	0.028057	0.302583	0.028010
Day	-0.061405	-0.051428	0.107706	-0.047261	0.013748	0.031227	1.000000	NaN	NaN	-0.006650	0.022738	0.105154	-0.003024
Month	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Dep_hour	-0.012587	0.023931	-0.075185	0.014079	-0.051677	-0.032077	-0.006650	NaN	NaN	1.000000	-0.004168	0.116047	0.021883
Dep_min	0.216567	-0.071648	0.031035	0.057221	-0.003191	0.028057	0.022738	NaN	NaN	-0.004168	1.000000	-0.003120	0.055340
Duration_hours	-0.004011	-0.037815	0.082008	-0.293599	0.022149	0.302583	0.105154	NaN	NaN	0.116047	-0.003120	1.000000	-0.014772
Duration_mins	-0.026288	0.008952	-0.059704	-0.005125	0.036267	0.028010	-0.003024	NaN	NaN	0.021883	0.055340	-0.014772	1.000000

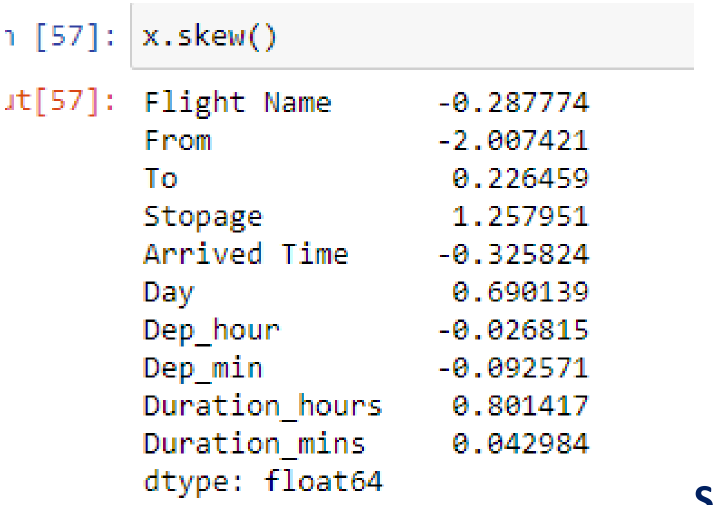
So , Here we can see that Month and Year have same value in all rows . So I can decided to delete this , and move foreword.

- Lets Check Correlation with Target Variable . With Barplot.



So , Here we can see that Duration Minute have Positve Relation. And Stopage have Some Negative Relationship.

- ➢ Then we Separate Features and Label .and then Check For skewness .



```
n [57]: x.skew()

ut[57]: Flight Name       -0.287774
        From              -2.007421
        To                 0.226459
        Stopage            1.257951
        Arrived Time      -0.325824
        Day                0.690139
        Dep_hour          -0.026815
        Dep_min           -0.092571
        Duration_hours     0.801417
        Duration_mins      0.042984
        dtype: float64
```

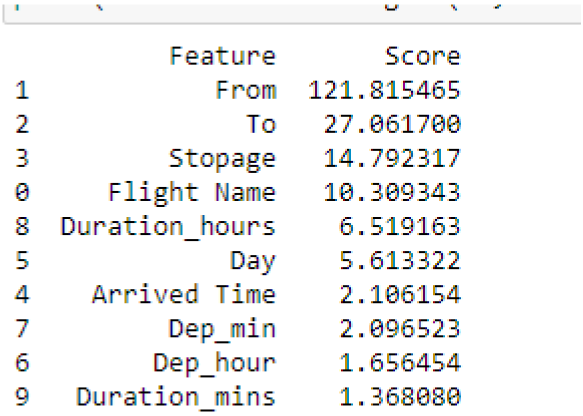
So , If we see only in Numerical Column we can find that we have some skewness in our features. Lets treat by power transformer.

- Hardware and Software Requirements and Tools Used
  - ➢ Anaconda Navigator -> Jupyter Notebook
  - ➢ Windows 11 - > Home Version
  - ➢ Hardware -> AMD Ryzen 3 Processor with Vega Graphics 2200 U.
  - ➢ Using Selenium Web Driver to Scrape the data from Yatra.com.

Model/s Development and Evaluation

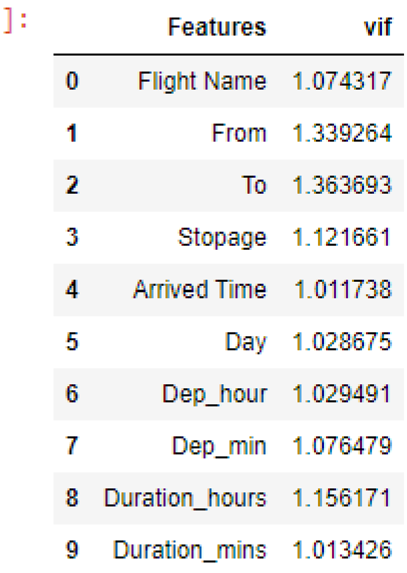
Here I am Separating the Feature and Label in X and Y . X is use for feature and Y is used for label . Then we check for skewness and we find some of the numerical columns have skewness. Then we resolve the skewness using power transformer . and we all know if we use power transformer then we don't need to use Standard Scaler .Standard Scaler Basically Normalize the data .

After that we use SelectKBest for feature selection . and we can see that all feature contributiong for predicting the model.



```
          Feature       Score
1            From  121.815465
2              To   27.061700
3         Stopage   14.792317
0     Flight Name   10.309343
8  Duration_hours    6.519163
5             Day    5.613322
4    Arrived Time    2.106154
7         Dep_min    2.096523
6        Dep_hour    1.656454
9   Duration_mins    1.368080
```

The, We check for multi collinearity problem using VIF (Variane Inflation Factor). And we set threshold for +/- 10 .



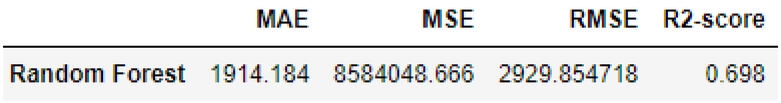
]:

	Features	vif
0	Flight Name	1.074317
1	From	1.339264
2	To	1.363693
3	Stopage	1.121661
4	Arrived Time	1.011738
5	Day	1.028675
6	Dep_hour	1.029491
7	Dep_min	1.076479
8	Duration_hours	1.156171
9	Duration_mins	1.013426

We can see that no columns have Multi collinearity Problem.

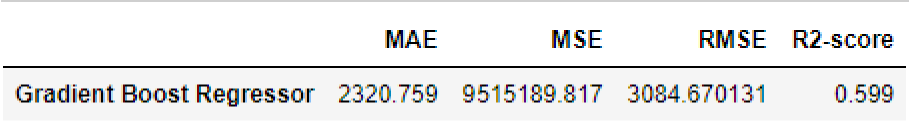
Lets Build a Model.

Random Forest ->



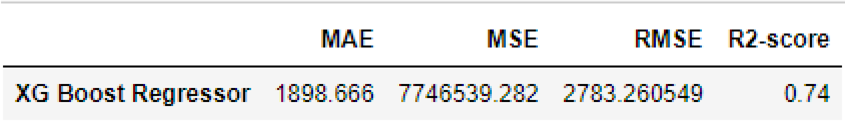
	MAE	MSE	RMSE	R2-score
Random Forest	1914.184	8584048.666	2929.854718	0.698

Gradient Boost ->



	MAE	MSE	RMSE	R2-score
Gradient Boost Regressor	2320.759	9515189.817	3084.670131	0.599

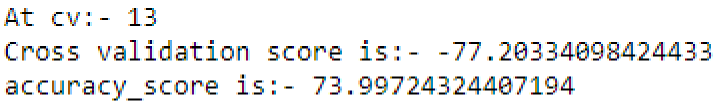
XGB ->



	MAE	MSE	RMSE	R2-score
XG Boost Regressor	1898.666	7746539.282	2783.260549	0.74

And , In Model Building time I am using almost all algorithm which I was learned . But I can show only three algorithm . and I am selecting XGB for model building because its gives a best accuracy .

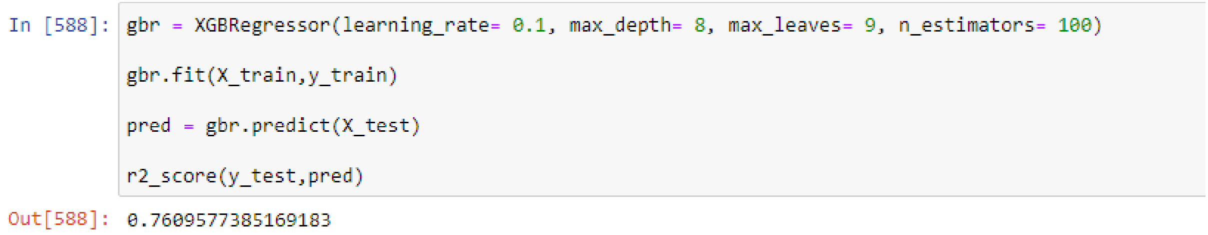
And, The Cross Validation Score ->



```
At cv:- 13
Cross validation score is:- -77.20334098424433
accuracy_score is:- 73.99724324407194
```

Lets Do Hyperparameter Tuning ->

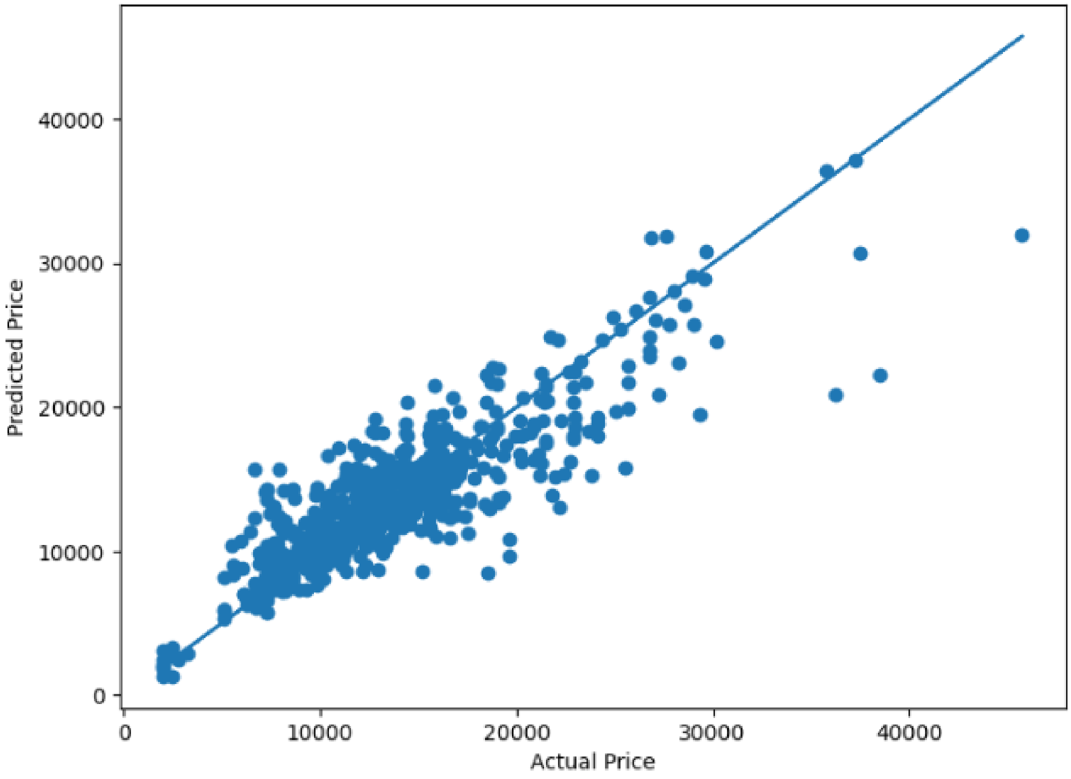
Using GridSearchCV ->



```
In [588]: gbr = XGBRegressor(learning_rate= 0.1, max_depth= 8, max_leaves= 9, n_estimators= 100)

          gbr.fit(X_train,y_train)

          pred = gbr.predict(X_test)

          r2_score(y_test,pred)

Out[588]: 0.7609577385169183
```

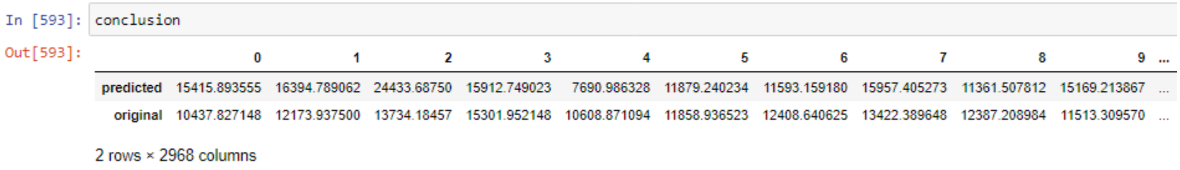
After , Tune the Parameter we get accuracy 76 % and our Cross Val Score is – 77 . So its means is our model is not overfitted .

And Now we are good to plot Best Fit Line .



Lets , See in the best fit line . we can clearly see that most of the data points comes near the best fit line . its means is our model is performing good . and we are go in the production.

CONCLUSION



```
In [593]: conclusion
Out[593]:
```

	0	1	2	3	4	5	6	7	8	9	...
predicted	15415.893555	16394.789062	24433.68750	15912.749023	7690.986328	11879.240234	11593.159180	15957.405273	11361.507812	15169.213867	...
original	10437.827148	12173.937500	13734.18457	15301.952148	10608.871094	11858.936523	12408.640625	13422.389648	12387.208984	11513.309570	...

2 rows × 2968 columns

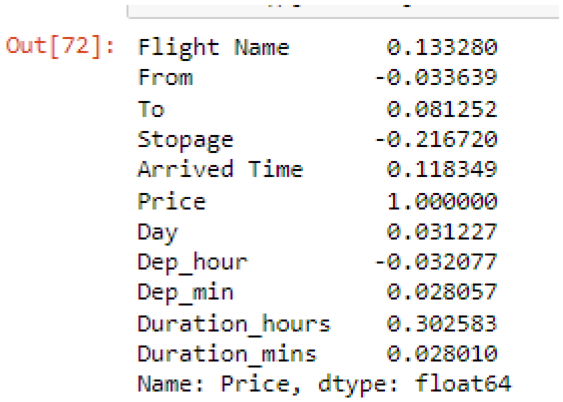
- Key Findings and Conclusions of the Study
  - ➢ Price of instant Ticket Booking is too much High
  - ➢ Price for 1 week late booking is low compare to instant booking
  - ➢ Most of the Flights comes from Vistara Airline then Indigo.
  - ➢ Most of flights are fly from New Delhi to Mumbai and Benglore
  - ➢ Most of the Flights are Non – Stop Flying . Direct land to the destination.

- Learning Outcomes of the Study in respect of Data Science..

  In the scraping time I can only scrap 9 Colums including price . and we have to predict the price . using machine learning algorithm and Coding language we use python .and for visualization we use seaborn and matplotlib .

and here. Price is Continious number so we have to use Regression algorithms.

Here we have very less columns and all the columns are very important for us to predict price . and we got all columns in object data type. So we converted Date in pd.datetime , and Price in Float data type ..after converting required columns. Then separate the date month and year. So we can find that month and year column have same value so I decided to delete. After deleting this . I am moving foreword to Checking correlation with Target column.



```
Out[72]: Flight Name        0.133280
         From              -0.033639
         To                 0.081252
         Stopage           -0.216720
         Arrived Time       0.118349
         Price              1.000000
         Day                0.031227
         Dep_hour          -0.032077
         Dep_min            0.028057
         Duration_hours     0.302583
         Duration_mins      0.028010
         Name: Price, dtype: float64
```

So , Here we can see that most of the relation with target comes from Flight name then Stopage then showing in above photo.

So, after all preprocessing when I started model building I got an issue is to with random forest and Gradient Boost . I got accracy is good. But Cross_Val_Score is too low . so I tried with removing outliers then I treid with removing less correlated features . but the Cross Validation Score is not getting high.

But when I try with Xtreme Gradient Boost I got good Accuracy and Good Cross Val Score in range of 15 .

Main Problem is facing due to model building to convert Arrived time in Hours .

- Limitations of this work and Scope for Future Work.

So , here we can scrap only next 5-6 days data . and we got 76 % accuracy . it means for next week flight price we predict .and if we use this Model for prediction we get good output for next week .

Limitations of this project is we have less number of features. If we get interior column, where we will get feature like, food etc. More the number of features, more accuracy we'll get .

In future, if someone do the proper and detail study of this dataset's each column than the accuracy will be so high.

THANKYOU