PRICE PREDICTION

Submitted by:

Prabhat Chauhan

ACKNOWLEDGMENT

I would like to mention my special thanks to flip robo for constantly providing us the opportunity to work on this projects which helped me to

learn new things, and I also pay my sincere thanks to Keshav Bansal sir for providing the support that is needed in between the projects, with his guidance I learned a lot and will continue to do so.

INTRODUCTION

• Flights are one of the necessary need of each and every person around the globe and therefore flight tickets market is the market which is one of the major contributors in the world's economy. It is a very large market and there are various companies working in the domain.

• Data science comes as a very important tool to solve problems in the domain to help the companies increase their overall revenue, profits, improving their marketing strategies and focusing on changing trends in flight tickets sales and purchases. Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques used for achieving the business goals for flight tickets selling companies. Our problem is related to one such flight-tickets selling client.

• We are required to model the price of tickets with the available independent variables. This model will then be used by the management to understand how exactly the prices vary with the variables. They can accordingly manipulate the strategy of the firm and concentrate on areas that will yield high returns. Further, the model will be a good way for the management to understand the pricing dynamics of a new market.

• With the covid 19 impact in the market, we have seen lot of changes in the flight tickets market. Now some flights are in demand hence making them costly and some are not in demand hence cheaper. One of the clients works with small traders, who sell flight tickets. With the change in market due to covid 19 impact, client is facing problems with their previous flight tickets price valuation machine learning models. So, they are looking for new machine learning models from new data.

• For this client wants to know:

➢ Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time?

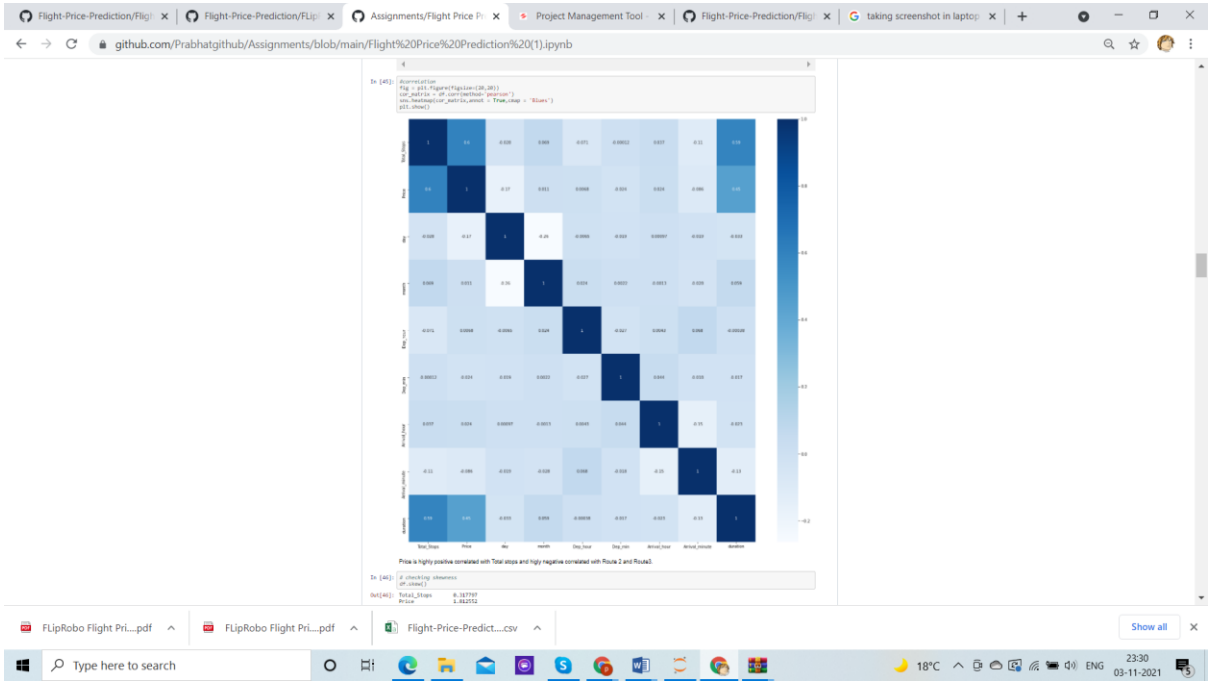
➢ What is the best time to buy so that the consumer can save the most by taking the least risk?

➢ Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

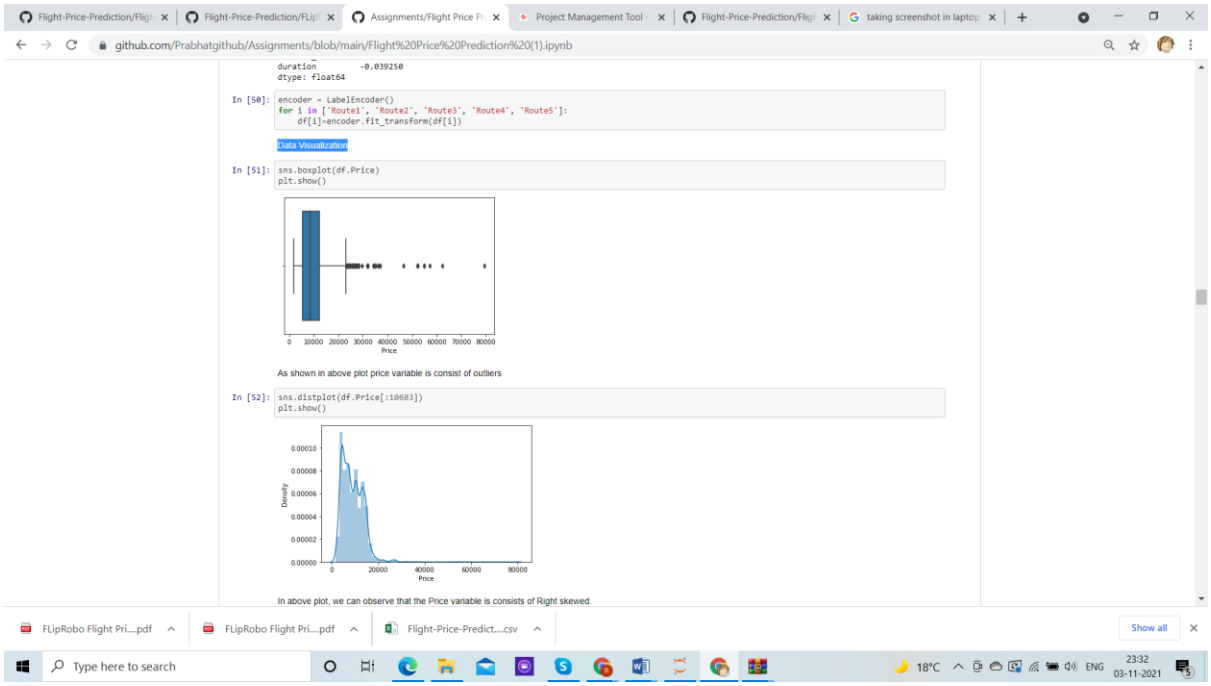
Analytical Problem Framing

- **Mathematical/ Analytical Modeling of the Problem**

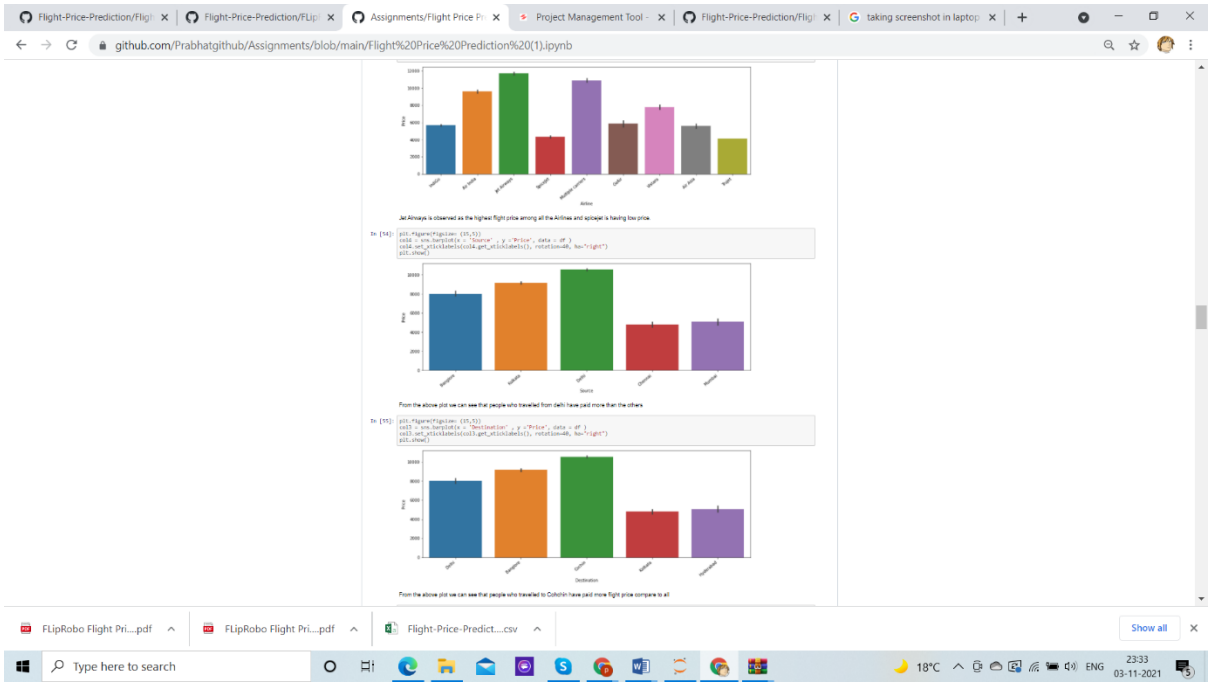
In this project we have performed various mathematical and statistical analysis such as we checked description or statistical summary of the data using describe, checked correlation using corr and also visualized it using heatmap. Then we have used zscore to plot outliers and remove them. Removing the Outliers using Z-score • Data Sources and their formats The sample data is extracted by using web scraping

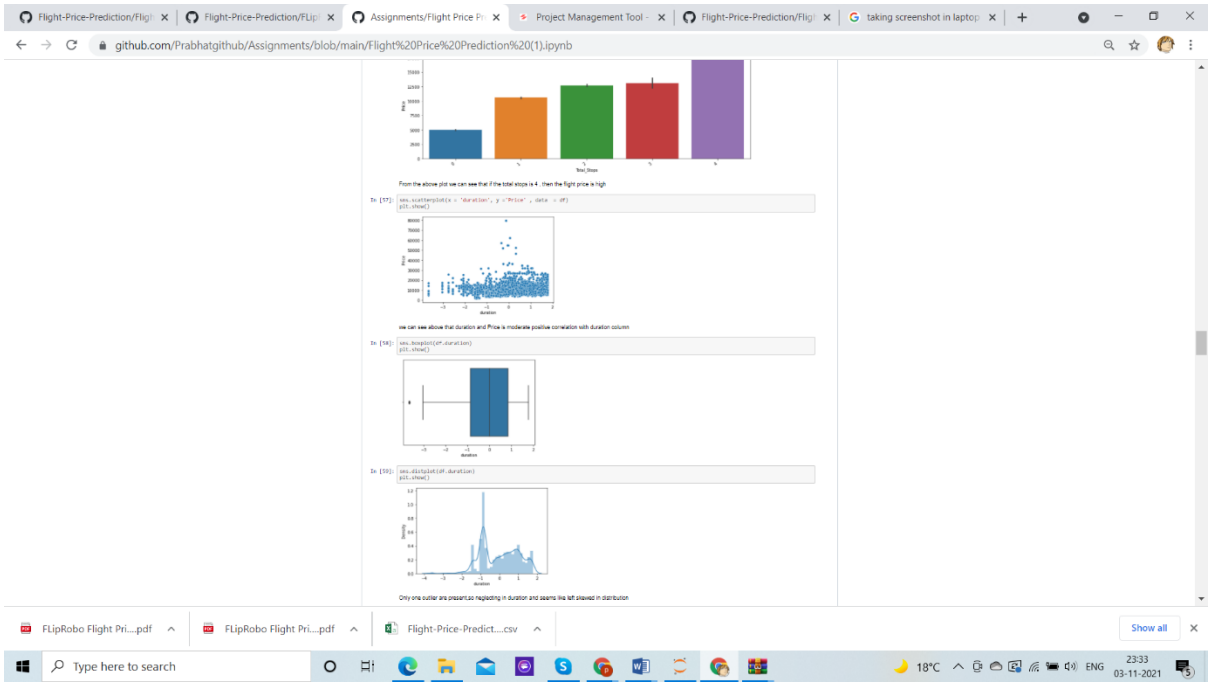


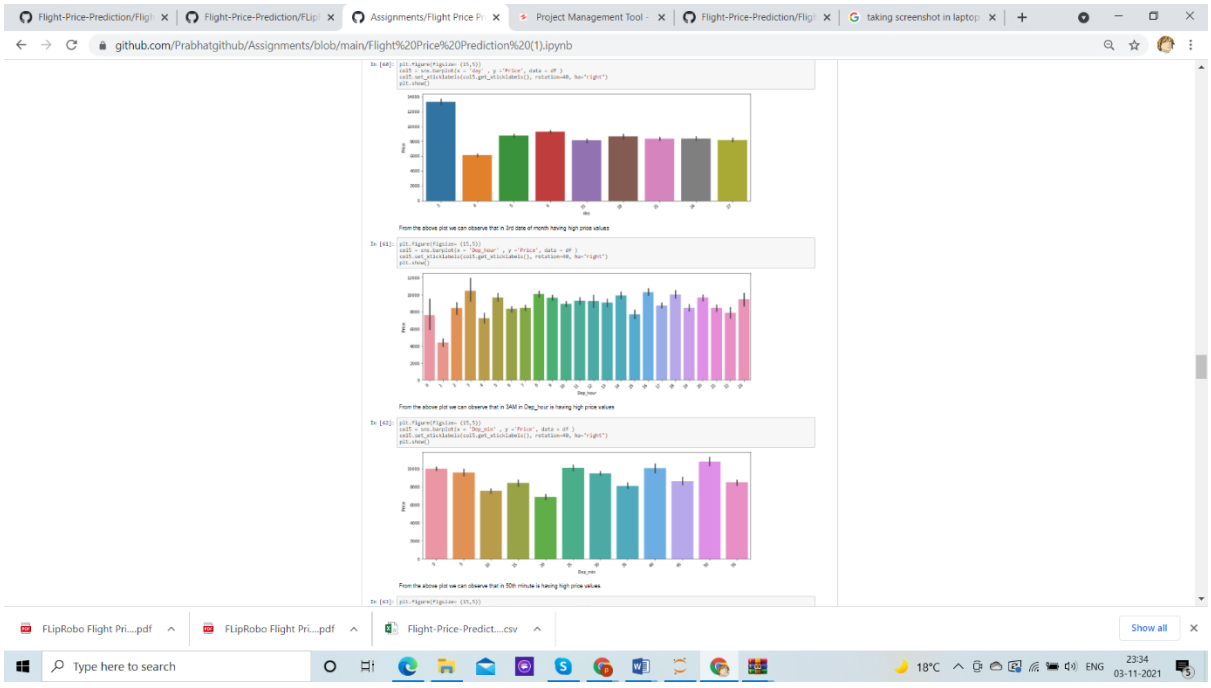
Data Visualization

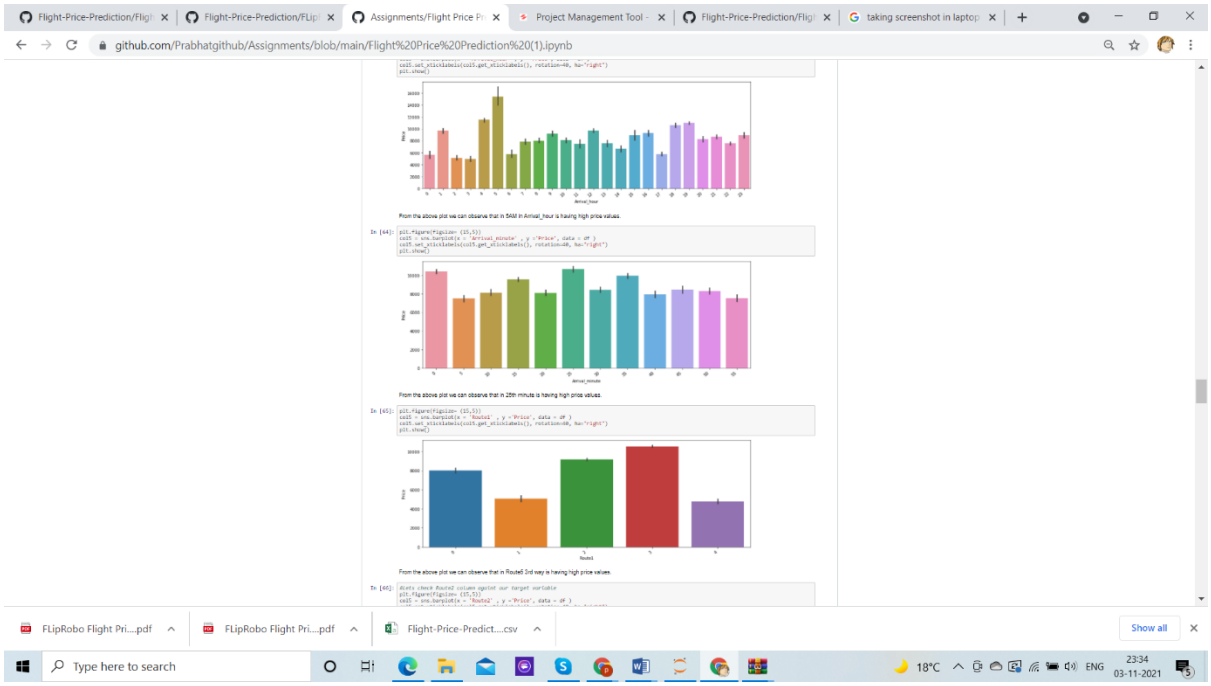


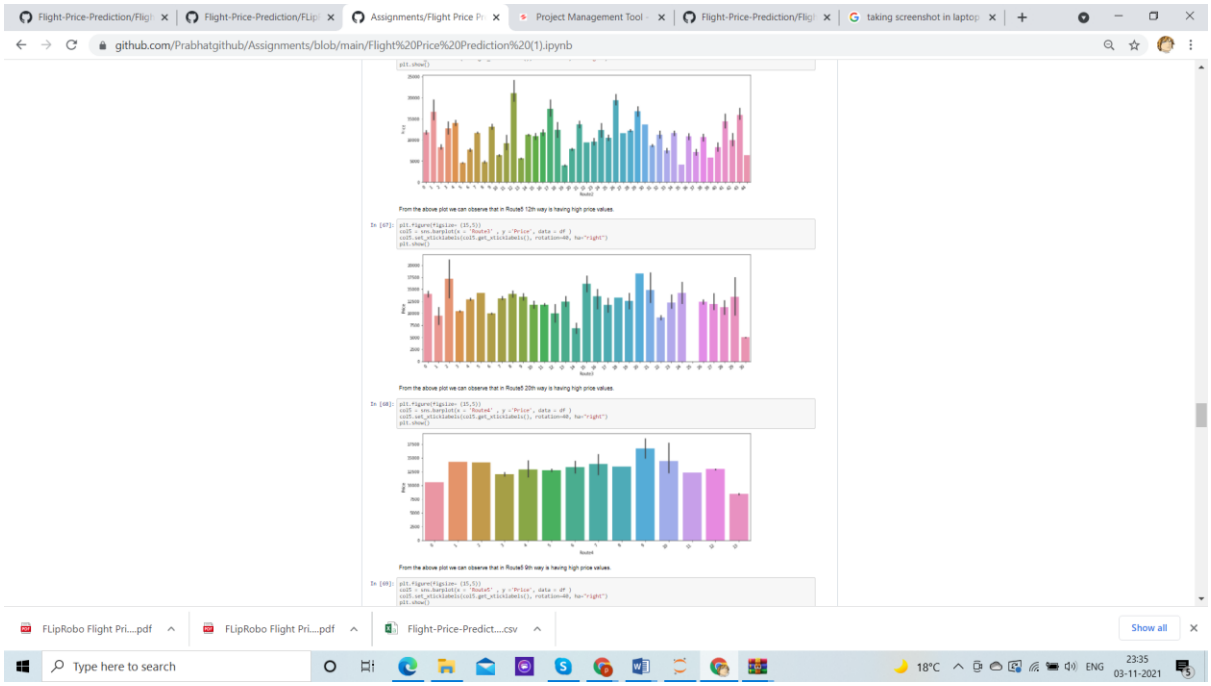
Bivariate Analysis




Flight-Price-Prediction/Flig
Flight-Price-Prediction/FLip
Assignments/Flight Price Pr
Project Management Tool -
Flight-Price-Prediction/Flig
taking screenshot in laptop
github.com/Prabhatgithub/Assignments/blob/main/Flight%20Price%20Prediction%20(1).ipynb
From the above plot we can see that if the total stops is 4 , then the flight price is high
In [57]:
plt.show()
we can see above that duration and Price is moderate positive correlation with duration column
In [58]:
plt.show()
In [59]:
plt.show()
Only one outlier are present,so neglecting in duration and seems like left skewed in distribution
FLipRobo Flight Pri....pdf
FLipRobo Flight Pri....pdf
Flight-Price-Predict....csv
Show all
Type here to search
18°C
ENG
23:33
03-11-2021


Flight-Price-Prediction/Flig
Flight-Price-Prediction/FLip
Assignments/Flight Price Pr
Project Management Tool -
Flight-Price-Prediction/Flig
taking screenshot in laptop
github.com/Prabhatgithub/Assignments/blob/main/Flight%20Price%20Prediction%20(1).ipynb
From the above plot we can observe that in 3rd date of month having high price values
In [61]:
From the above plot we can observe that in 3AM in Dep_hour is having high price values
In [62]:
From the above plot we can observe that in 50th minute is having high price values.
In [63]:
FLipRobo Flight Pri....pdf
FLipRobo Flight Pri....pdf
Flight-Price-Predict....csv
Show all
Type here to search
18°C
ENG
23:34
03-11-2021





Outlier Removal in target column

Flight-Price-Prediction/Fligh | Flight-Price-Prediction/FLip | Assignments/Flight Price Pr | Project Management Tool | Flight-Price-Prediction/Fligh | taking screenshot in laptop

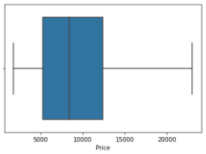
github.com/Prabhatgithub/Assignments/blob/main/Flight%20Price%20Prediction%20(1).ipynb

In [70]:
```
def outliers_IQR3(data_frame,feature_name):
    IQR =data_frame[feature_name].quantile(0.75)-data_frame[feature_name].quantile(0.25)
    lower_boundary = data_frame[feature_name].quantile(0.25) - (IQR * 1.5)
    upper_boundary = data_frame[feature_name].quantile(0.75) + (IQR * 1.5)
    df.loc[df[feature_name] >= upper_boundary,feature_name]= upper_boundary
    df.loc[df[feature_name] <= lower_boundary,feature_name]= lower_boundary
```

In [71]:
```
#calling the function
outliers_IQR3(df,'Price')
```

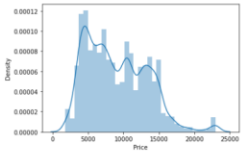
In [72]:
```
#now lets outliers using box plot
sns.boxplot(df.Price)
```

Out[72]: <AxesSubplot:xlabel='Price'>



outliers have been completely removed from this column

In [73]:
```
#lets check the distribution of the target column
sns.distplot(df.Price[:10683])
plt.show()
```



FLipRobo Flight Pri....pdf | FLipRobo Flight Pri....pdf | Flight-Price-Predict....csv | Show all

Type here to search | 18°C | ENG | 23:36 | 03-11-2021

Flight-Price-Prediction/Fligh | Flight-Price-Prediction/FLip | Assignments/Flight Price Pr | Project Management Tool | Flight-Price-Prediction/Fligh | taking screenshot in laptop

github.com/Prabhatgithub/Assignments/blob/main/Flight%20Price%20Prediction%20(1).ipynb

In [104]:
```
model_list = [dtc,knn,rf,ada]
least_difference = []
for m in model_list:
    m.fit(x_train,y_train)
    pred = m.predict(x_test)
    cvs = cross_val_score(m,X_train_original,Y,cv =5)
    print('\n')
    print(m)
    print('Scores :')
    print('r2 score: ',r2_score(y_test,pred))
    print('Cross Val score :',cvs.mean())
    print('Error :')
    print('mean absolute error :',mean_absolute_error(y_test,pred))
    print('mean squared error :', mean_squared_error(y_test,pred))
    print('root mean squared error :',np.sqrt(mean_squared_error(y_test,pred)))
    print('Difference :')
    difference = np.abs(r2_score(y_test,pred) - cvs.mean())
    print('Diffrence between cross val score and r2 score is : {0:.2f}'.format(difference))
    least_difference.append((m,'Diffrence between cross val score and r2 score error is : {0:.2f}'.format(difference)))
```

```
DecisionTreeRegressor()
Scores :
r2 score: 0.6488469462334607
Cross Val score : 0.661718522239223
Error :
mean absolute error : 1369.257337303275
mean squared error :  6390604.628243301
root mean squared error : 2527.96452274222
Difference :
Diffrence between cross val score and r2 score is : 0.01


KNeighborsRegressor()
Scores :
r2 score: 0.8090013530 1018
Cross Val score : 0.7934614014133123
Error :
mean absolute error : 1175.7146746065505
mean squared error :  3475968.169289664
root mean squared error : 1864.3948533745913
Difference :
Diffrence between cross val score and r2 score is : 0.02


RandomForestRegressor()
Scores :
r2 score: 0.8358978080351679
Cross Val score : 0.8201448481717607
Error :
mean absolute error : 1101.4279120975202
mean squared error :  2986481.8666009754
root mean squared error : 1728.1440526185818
Difference :
Diffrence between cross val score and r2 score is : 0.02


AdaBoostRegressor()
```

FLipRobo Flight Pri....pdf | FLipRobo Flight Pri....pdf | Flight-Price-Predict....csv | Show all

Type here to search | 18°C | ENG | 23:38 | 03-11-2021