FLIP ROBO

FLIGHT PRICE PREDICTION PROJECT

Submitted by:

CHAMLIN NAJIR RAHMAN

BATCH No. 1838

Internship 25
2022

ACKNOWLEDGMENT

I would like to express my special gratitude to “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analyzation skills. And I want to express my huge gratitude to Mr. Mohd Kashif (SME Flip Robo), he is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to my academic team “Datatrained” who are the reason behind what I am today. Last but not least my parents who have been my backbone in every step of my life. And also thank you for many other persons who has helped me directly or indirectly to complete the project.

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

• Mathematical/ Analytical Modeling of the Problem

As a first step I have scrapped the required data from yatra.com website. I have fetched data for different locations and saved it to excel format.

In this particular problem I have price as my target column and it was a continuous column. So clearly it is a regression problem and I have to use all regression algorithms while building the model. There were no null values in the dataset. Since we have scrapped the data from yatra.com website the raw data was not in the format, so we have use feature engineering to extract the required feature format. To get better insight on the features I have used plotting like distribution plot, bar plot, line plot, strip plot and count plot. With these plotting I was able to understand the relation between the features in better manner. Also, I found outliers and skewness in the dataset so I removed outliers using z-score method and I removed skewness using yeo-johnson method. I have used all the regression algorithms while building model then tuned the best model and saved the best model. At last I have predicted the car-price using saved model.

• Data Sources and their formats

The data was collected from yatra.com website in excel format. The data was scrapped using selenium. After scrapping required features the dataset is saved as excel file.

Also, my dataset was having 6730 rows and 8 columns including target. In this particular datasets I have object type of data which has been changed as per our analysis about the dataset.

• Data Preprocessing Done

- ✓ As a first step I have scrapped the required data using selenium from yatra website.
- ✓ And I have imported required libraries and I have imported the dataset which was in excel format.
- ✓ Then I did all the statistical analysis like checking shape, nunique, value counts, info etc…..

- ✓ While checking for null values I found there were no null values in the dataset.
- ✓ Next as a part of feature extraction I converted the data types of all the columns and I have extracted useful information from the raw dataset. Thinking that this data will help us more than raw data.
- ✓ I have also dropped Airline, Dep_Time, Arrival_Time, Duration, as I have cleaned them and extracted the required information from them.

• Data Inputs- Logic- Output Relationships

Since I had both categorical and numerical columns I have plotted dist plot to see the distribution of skewness in each column data and countplot for categorical .

- ➢ I have used bar plot for each pair of categorical features that shows the relation between label and independent features.
- ➢ I have used line plot and strip plot to see the relation between numerical columns with target column.
- ➢ I can notice there is a linear relationship between maximum columns and target

• Hardware and Software Requirements and Tools Used

While taking up the project we should be familiar with the Hardware and software required for the successful completion of the project. Here we need the following hardware and software.

Hardware required: -

1. Processor — core i5 and above
2. RAM — 8 GB or above
3. SSD — 250GB or above

Software/s required: -

1. Anaconda
2. Jupyter

Model/s Development and Evaluation

- **Identification of possible problem-solving approaches (methods)**

  Since the data collected was not in the format we have to clean it and bring it to the proper format for our analysis. To remove outliers I have used z-score method. And to remove skewness I have used yeo-johnson method. We have dropped all the unnecessary columns in the dataset according to our understanding. Use of Pearson's correlation coefficient to check the correlation between dependent and independent features. Also I have used Standardisation to scale the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms.

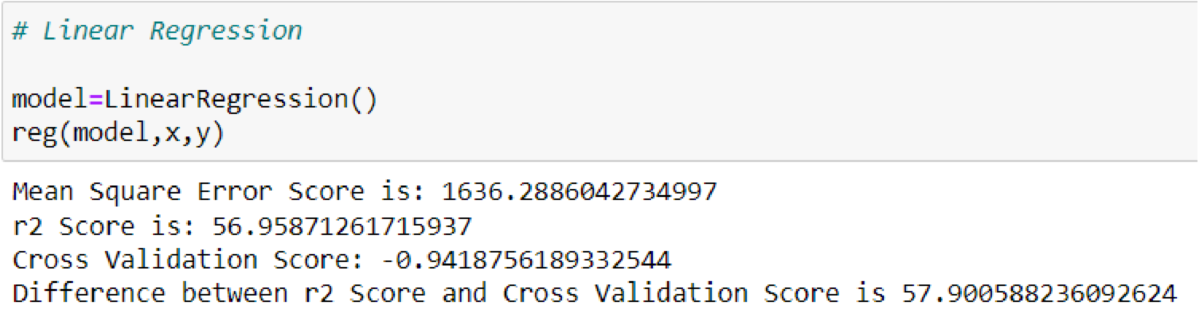
- **Testing of Identified Approaches (Algorithms)**

  Since price was my target and it was a continuous column with improper format which has to be changed to continuous float datatype column, so this perticular problem was Regression problem. And I have used all Regression algorithms to build my model. By looking into the difference of r2 score and cross validation score I found GradientBoostingRegressor as a best model with least difference. Also to get the best model we have to run through multiple models and to avoid the confusion of overfitting we have gone through cross validation. Below are the list of Regression algorithms I have used in my project.

- RandomForestRegressor
- Linear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet
- Support Vector Regression (poly)
- Support Vector Regression (linear)
- Support Vector Regression (rbf)

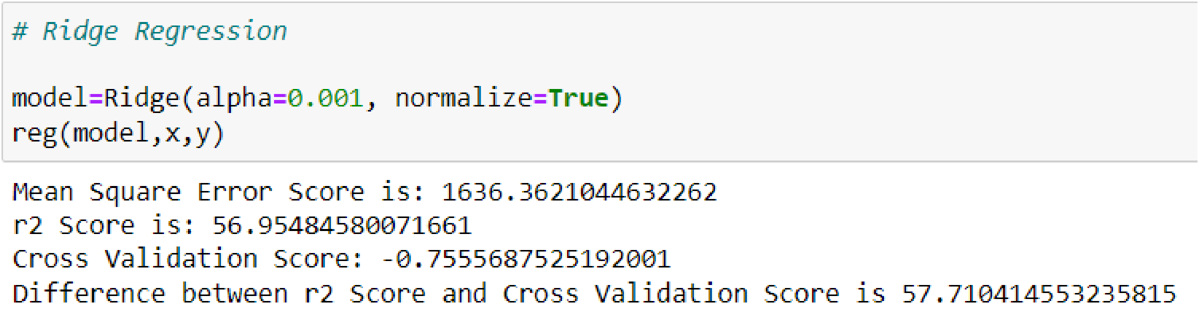
- K Neighbors Regressor
- ExtraTreesRegressor
- Gradient Boosting Regressor
- Decision Tree Regressor
- Ada Boost Regressor

- **Run and Evaluate selected models**
  All the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics :



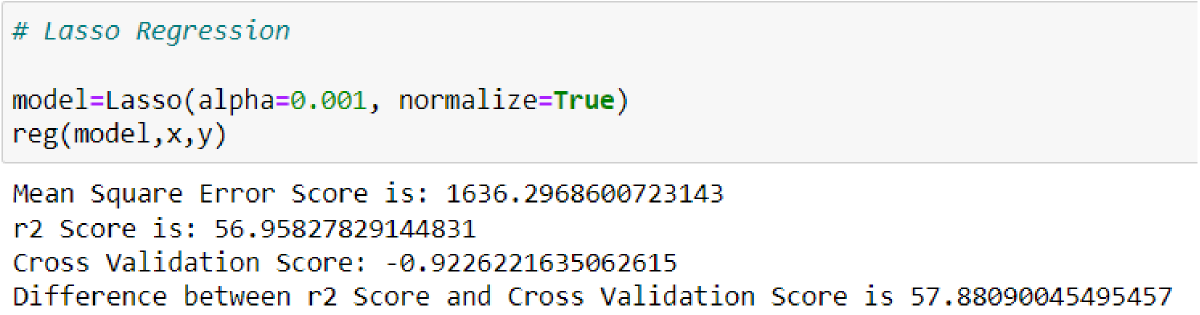
```
# Linear Regression

model=LinearRegression()
reg(model,x,y)
```

```
Mean Square Error Score is: 1636.2886042734997
r2 Score is: 56.95871261715937
Cross Validation Score: -0.9418756189332544
Difference between r2 Score and Cross Validation Score is 57.900588236092624
```



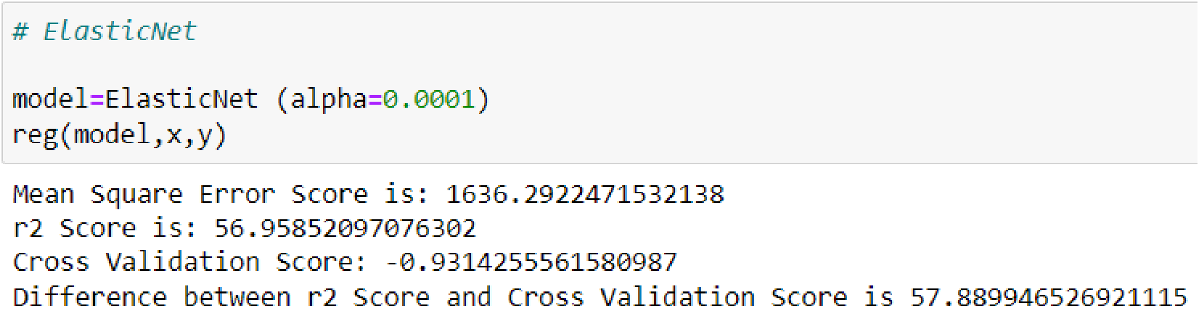
```
# Ridge Regression

model=Ridge(alpha=0.001, normalize=True)
reg(model,x,y)
```

```
Mean Square Error Score is: 1636.3621044632262
r2 Score is: 56.95484580071661
Cross Validation Score: -0.7555687525192001
Difference between r2 Score and Cross Validation Score is 57.710414553235815
```



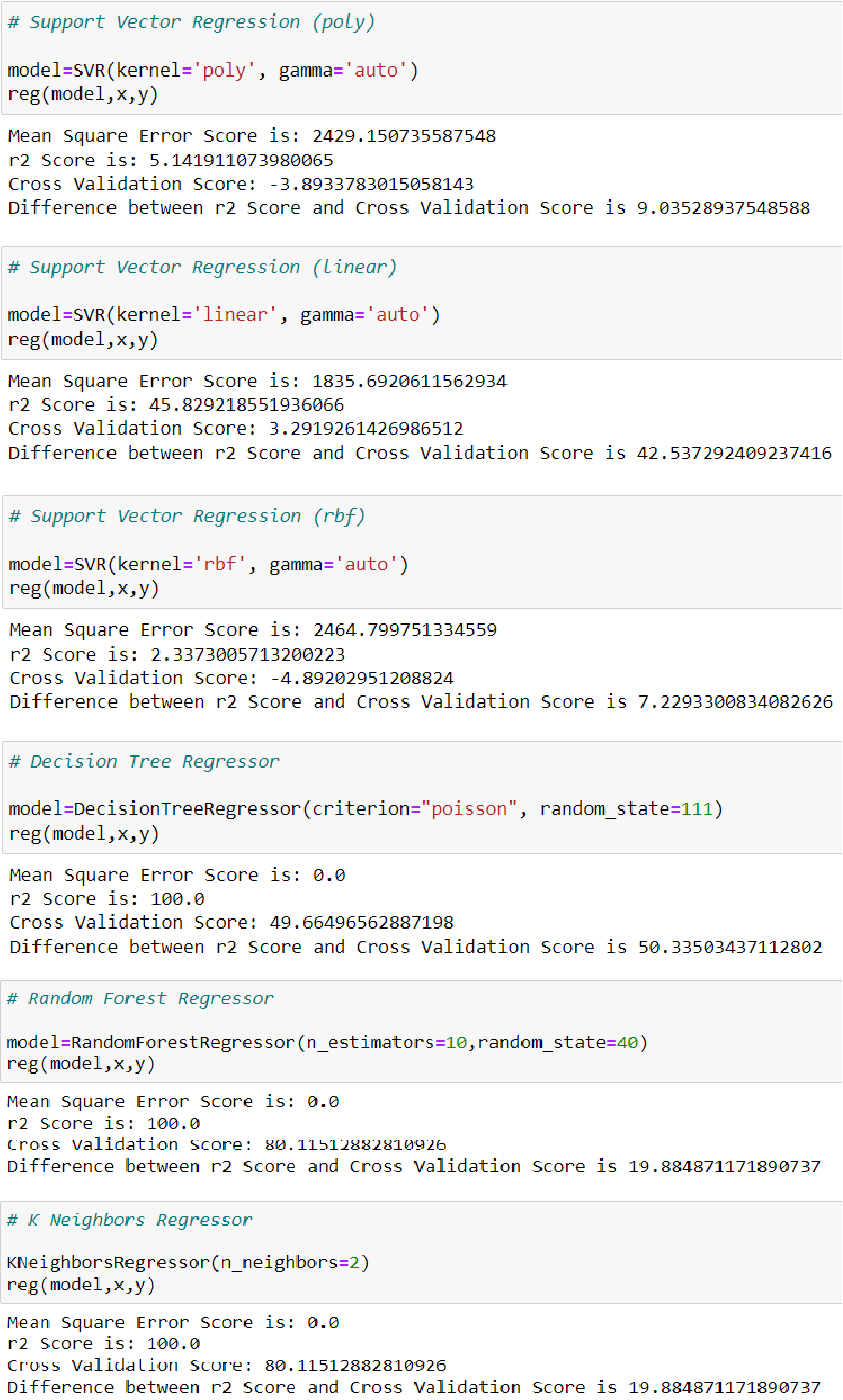
```
# Lasso Regression

model=Lasso(alpha=0.001, normalize=True)
reg(model,x,y)
```

```
Mean Square Error Score is: 1636.2968600723143
r2 Score is: 56.95827829144831
Cross Validation Score: -0.9226221635062615
Difference between r2 Score and Cross Validation Score is 57.88090045495457
```



```
# ElasticNet

model=ElasticNet (alpha=0.0001)
reg(model,x,y)
```

```
Mean Square Error Score is: 1636.2922471532138
r2 Score is: 56.95852097076302
Cross Validation Score: -0.9314255561580987
Difference between r2 Score and Cross Validation Score is 57.889946526921115
```



```
# Support Vector Regression (poly)

model=SVR(kernel='poly', gamma='auto')
reg(model,x,y)
```

```
Mean Square Error Score is: 2429.150735587548
r2 Score is: 5.141911073980065
Cross Validation Score: -3.8933783015058143
Difference between r2 Score and Cross Validation Score is 9.03528937548588
```

```
# Support Vector Regression (linear)

model=SVR(kernel='linear', gamma='auto')
reg(model,x,y)
```

```
Mean Square Error Score is: 1835.6920611562934
r2 Score is: 45.829218551936066
Cross Validation Score: 3.2919261426986512
Difference between r2 Score and Cross Validation Score is 42.537292409237416
```

```
# Support Vector Regression (rbf)

model=SVR(kernel='rbf', gamma='auto')
reg(model,x,y)
```

```
Mean Square Error Score is: 2464.799751334559
r2 Score is: 2.3373005713200223
Cross Validation Score: -4.89202951208824
Difference between r2 Score and Cross Validation Score is 7.2293300834082626
```

```
# Decision Tree Regressor

model=DecisionTreeRegressor(criterion="poisson", random_state=111)
reg(model,x,y)
```

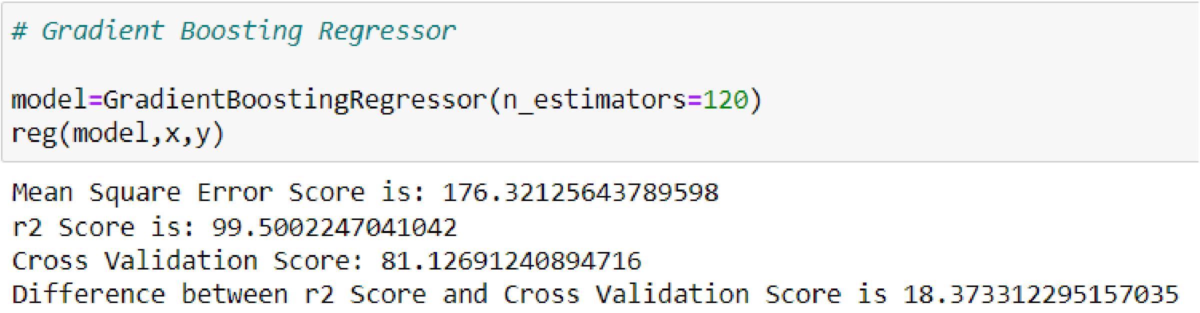
```
Mean Square Error Score is: 0.0
r2 Score is: 100.0
Cross Validation Score: 49.66496562887198
Difference between r2 Score and Cross Validation Score is 50.33503437112802
```

```
# Random Forest Regressor

model=RandomForestRegressor(n_estimators=10,random_state=40)
reg(model,x,y)
```

```
Mean Square Error Score is: 0.0
r2 Score is: 100.0
Cross Validation Score: 80.11512882810926
Difference between r2 Score and Cross Validation Score is 19.884871171890737
```

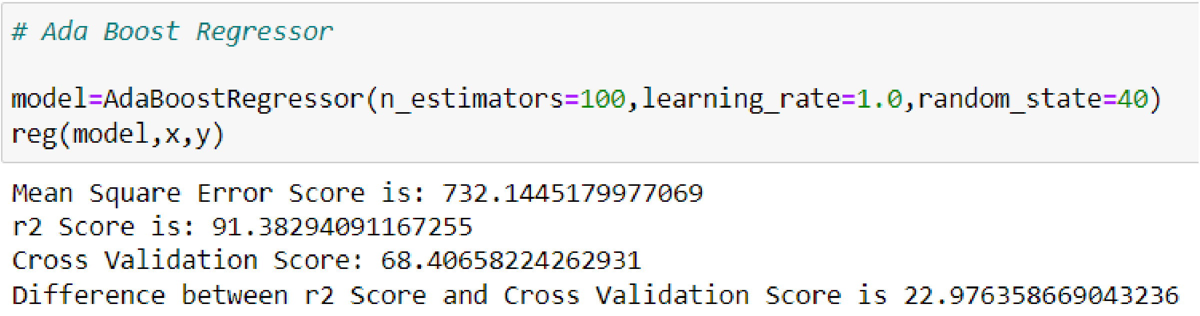
```
# K Neighbors Regressor

KNeighborsRegressor(n_neighbors=2)
reg(model,x,y)
```

```
Mean Square Error Score is: 0.0
r2 Score is: 100.0
Cross Validation Score: 80.11512882810926
Difference between r2 Score and Cross Validation Score is 19.884871171890737
```



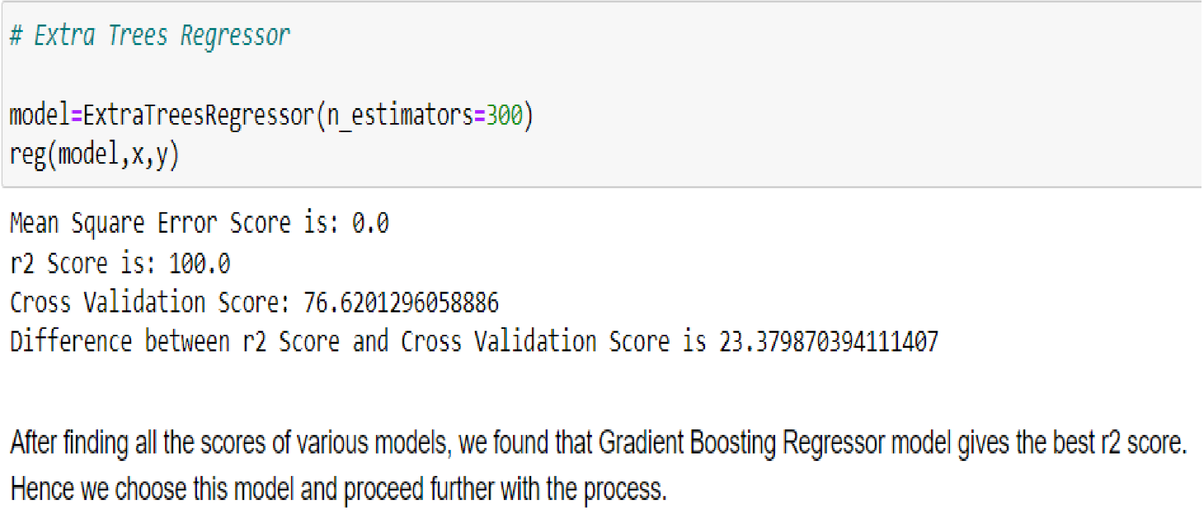
```
# Gradient Boosting Regressor

model=GradientBoostingRegressor(n_estimators=120)
reg(model,x,y)
```

```
Mean Square Error Score is: 176.32125643789598
r2 Score is: 99.5002247041042
Cross Validation Score: 81.12691240894716
Difference between r2 Score and Cross Validation Score is 18.373312295157035
```



```
# Ada Boost Regressor

model=AdaBoostRegressor(n_estimators=100,learning_rate=1.0,random_state=40)
reg(model,x,y)
```

```
Mean Square Error Score is: 732.1445179977069
r2 Score is: 91.38294091167255
Cross Validation Score: 68.40658224262931
Difference between r2 Score and Cross Validation Score is 22.976358669043236
```



```
# Extra Trees Regressor

model=ExtraTreesRegressor(n_estimators=300)
reg(model,x,y)
```

```
Mean Square Error Score is: 0.0
r2 Score is: 100.0
Cross Validation Score: 76.6201296058886
Difference between r2 Score and Cross Validation Score is 23.379870394111407
```

After finding all the scores of various models, we found that Gradient Boosting Regressor model gives the best r2 score. Hence we choose this model and proceed further with the process.

- ## Key Metrics for success in solving problem under consideration

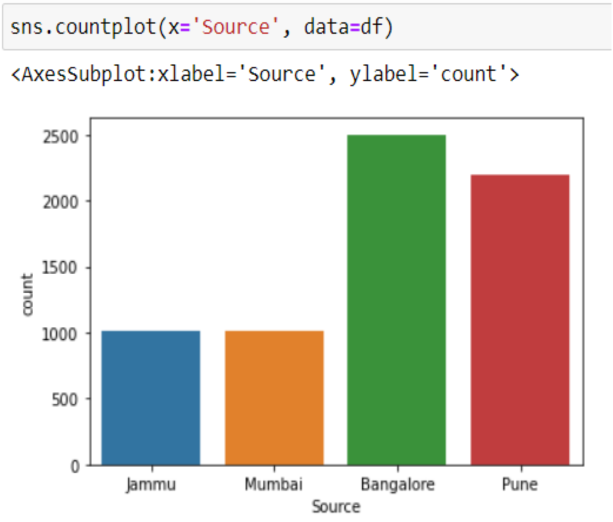
I have used the following metrics for evaluation:

- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
- I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
- I have used r2 score which tells us how accurate our model is.

- **Visualization**

  I have used bar plots to see the relation of categorical feature with target and I have used 2 types of plots for numerical columns one is disp plot for univariate and, strip plot for bivariate analysis.

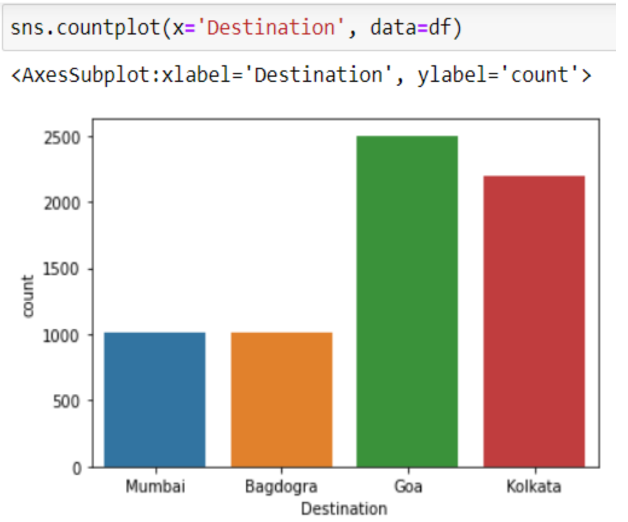
Univariate analysis for categorical column :



```
sns.countplot(x='Source', data=df)
```

```
<AxesSubplot:xlabel='Source', ylabel='count'>
```

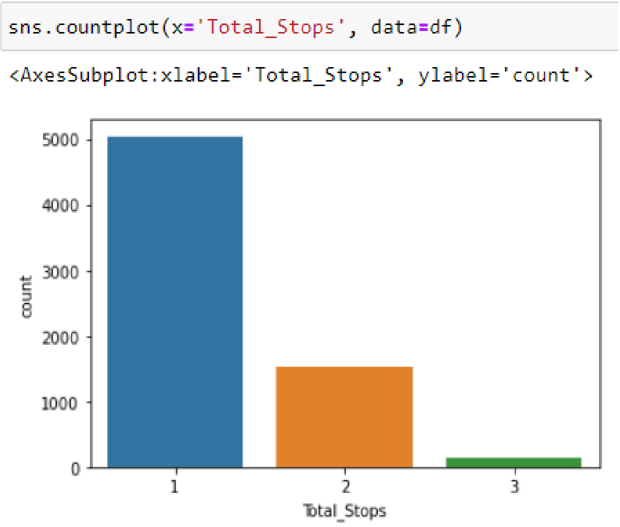
Mostly the people are travelling from bangalore.



```
sns.countplot(x='Destination', data=df)
```

```
<AxesSubplot:xlabel='Destination', ylabel='count'>
```

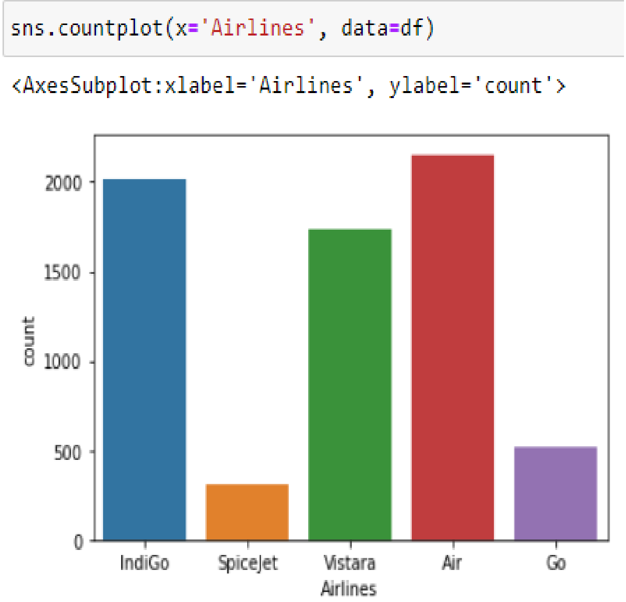
And the destinstion is mostly Goa and kolkata.



```
sns.countplot(x='Total_Stops', data=df)
```

```
<AxesSubplot:xlabel='Total_Stops', ylabel='count'>
```

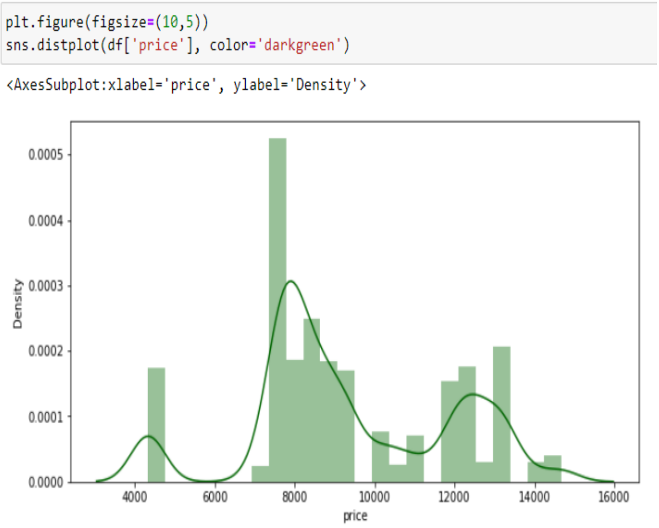
Most people usually prefer routes with 1 stopage.



```
sns.countplot(x='Airlines', data=df)
```

```
<AxesSubplot:xlabel='Airlines', ylabel='count'>
```

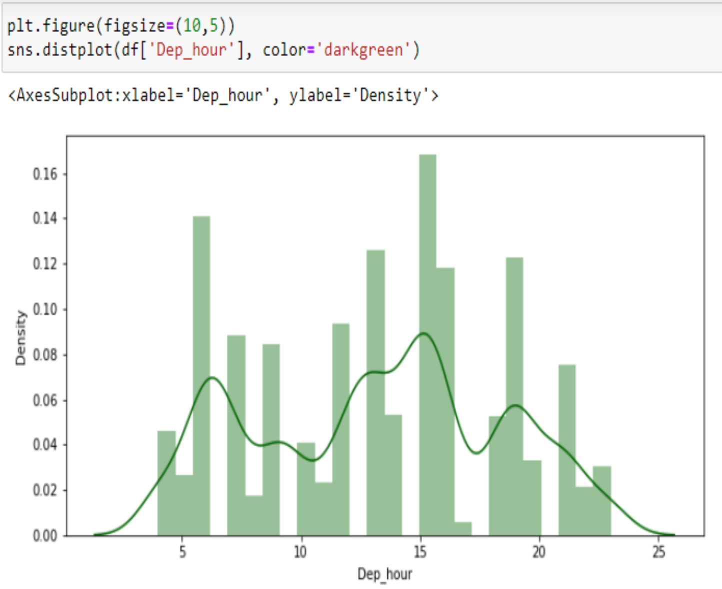
Univariate analysis for numerical column :



```
plt.figure(figsize=(10,5))
sns.distplot(df['price'], color='darkgreen')
```

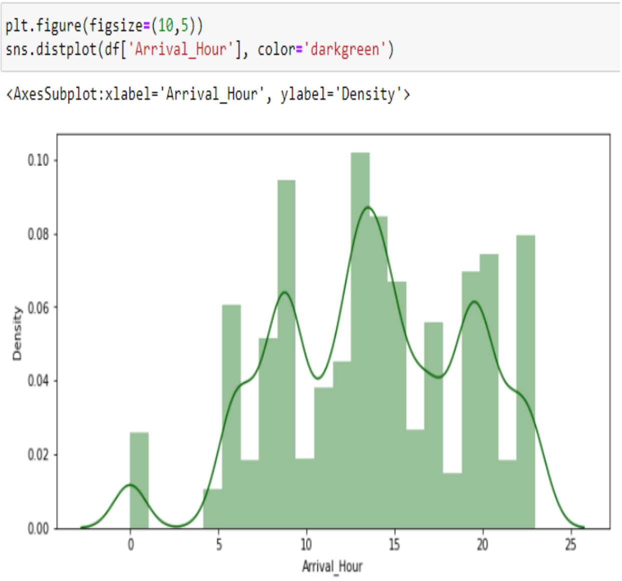
```
<AxesSubplot:xlabel='price', ylabel='Density'>
```

Mostly the flight prices ranges around Rs. 8000.



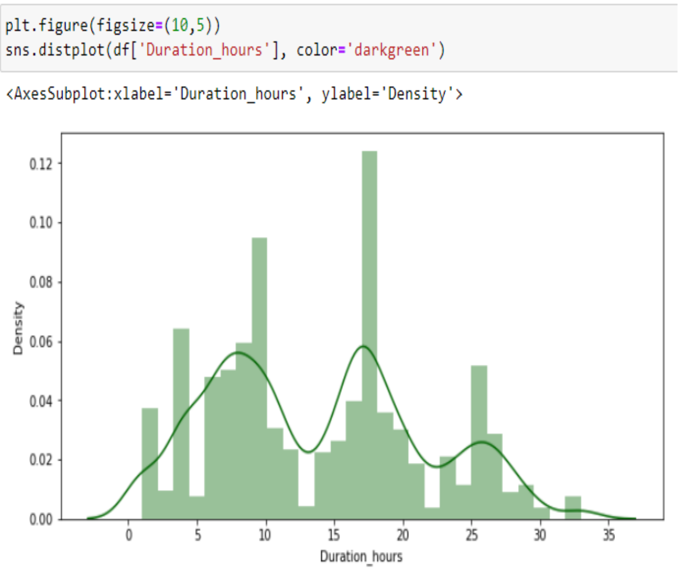
```
plt.figure(figsize=(10,5))
sns.distplot(df['Dep_hour'], color='darkgreen')
```

```
<AxesSubplot:xlabel='Dep_hour', ylabel='Density'>
```



```
plt.figure(figsize=(10,5))
sns.distplot(df['Arrival_Hour'], color='darkgreen')
```

```
<AxesSubplot:xlabel='Arrival_Hour', ylabel='Density'>
```



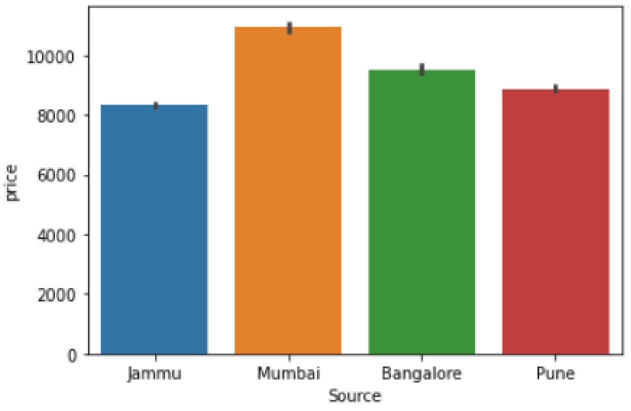
```
plt.figure(figsize=(10,5))
sns.distplot(df['Duration_hours'], color='darkgreen')
```

```
<AxesSubplot:xlabel='Duration_hours', ylabel='Density'>
```

Bivariate analysis for categorical column :

```
sns.barplot(data=df, x='Source',y='price')
```

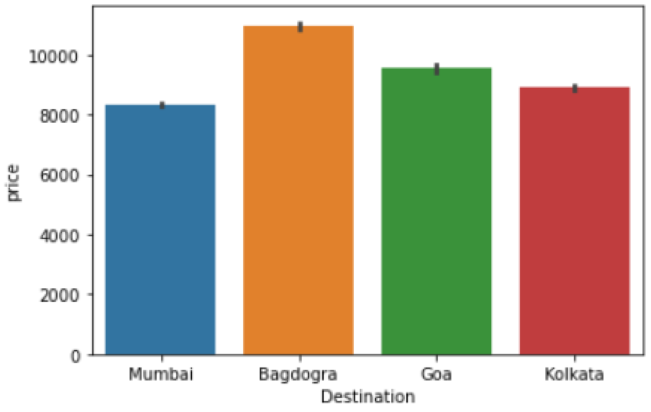
```
<AxesSubplot:xlabel='Source', ylabel='price'>
```



Flight price from mumbai is highest.

```
sns.barplot(data=df, x='Destination',y='price')
```

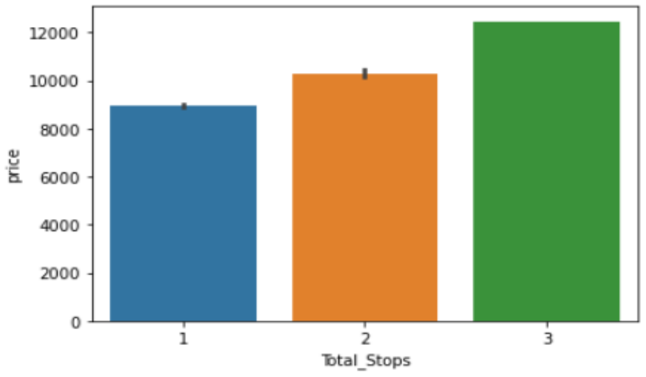
```
<AxesSubplot:xlabel='Destination', ylabel='price'>
```



Flight price for bagdogra is highest.

```
sns.barplot(data=df, x='Total_Stops',y='price')
```

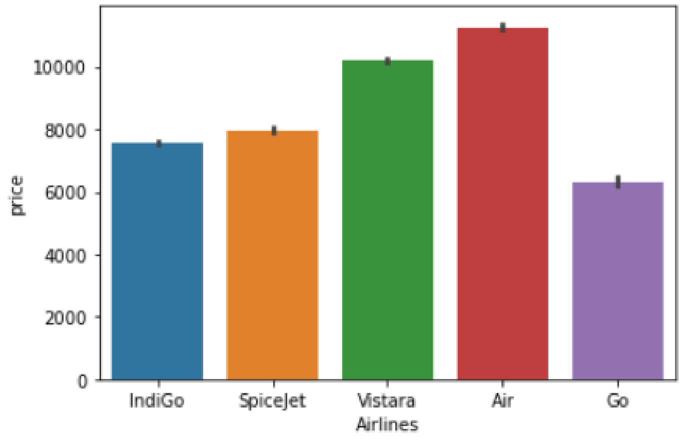
```
<AxesSubplot:xlabel='Total_Stops', ylabel='price'>
```



Flight prices are highest for routes with 3 stoppages.

```
sns.barplot(data=df, x='Airlines',y='price')
```

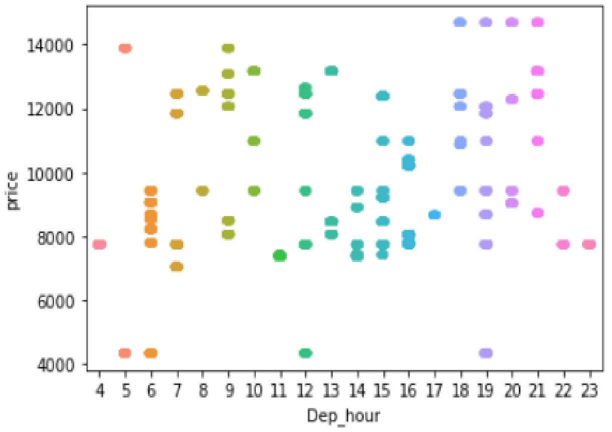
```
<AxesSubplot:xlabel='Airlines', ylabel='price'>
```



Bivariate analysis for numerical column :

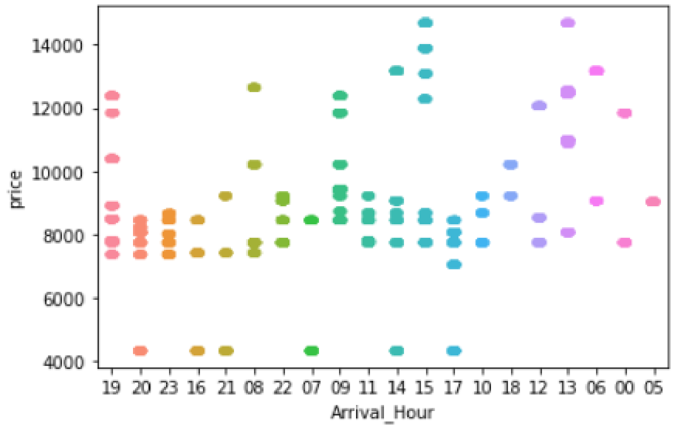
```
sns.stripplot(data=df, x='Dep_hour',y='price')
```

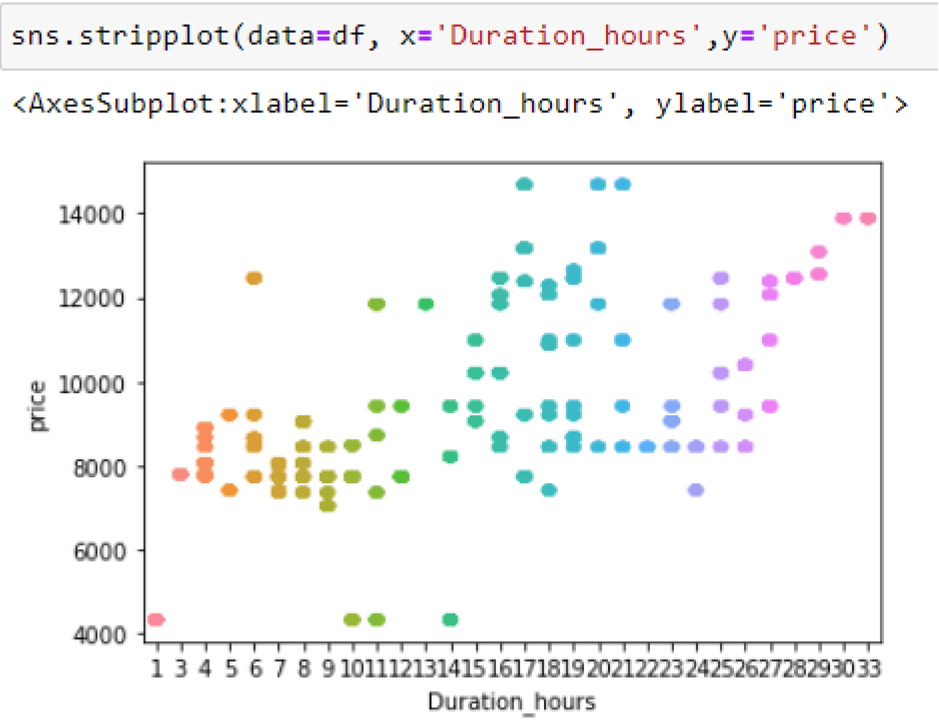
```
<AxesSubplot:xlabel='Dep_hour', ylabel='price'>
```



```
sns.stripplot(data=df, x='Arrival_Hour',y='price')
```

```
<AxesSubplot:xlabel='Arrival_Hour', ylabel='price'>
```





```
sns.stripplot(data=df, x='Duration_hours',y='price')
```

```
<AxesSubplot:xlabel='Duration_hours', ylabel='price'>
```

- **Interpretation of the Results**

- Data was having both numerical and continuous features.
- I did EDA to understand the data and written the observation also
- I did all the preprocessing from data cleaning to data transformation and also did Feature engineering
- Only 2% outliers was present.
- No skewness found.
- Finally i build a model that was giving me 99% CSV accuracy
- And finally my model can predict the flight price

CONCLUSION

• Key Findings and Conclusions of the Study

- Airfares do not change much frequently.
- They move in small increments.
- Yes, the flight prices tend to go up or down over time.
- The best time to buy tickets are over 1 week or more than that.
- Yes, prices increases as the departure data approaches.
- Morning flights are sometimes expensive, depends on the location.

• Learning Outcomes of the Study in respect of Data Science

In this project report, we have used machine learning algorithms to predict the used flight prices. We have mentioned the step by step procedure to analyze the dataset and finding the correlation between the features. Thus we can select the features which are correlated to each other and are independent in nature. These feature set were then given as an input to five algorithms and a hyper parameter tuning was done to the best model and the accuracy has been improved. Hence we calculated the performance of each model using different performance metrics and compared them based on those metrics. Then we have also saved the best model and predicted the flight price. It was good the predicted and actual values were almost same.

• Limitations of this work and Scope for Future Work

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.