
# PROJECT REPORT ON:

# “Flight Price Prediction”

## SUBMITTED BY

Adi Rajit Mahesh

# ACKNOWLEDGMENT

I would like to extend my deepest gratitude to “Flip Robo” team who have given me this opportunity to work on this amazing project. I would also like to thank “Srishti Mann” ma’am to extend helped me out during my project.

A huge thanks to “Datatrained” who are the reason for who I am today. Lastly, I would like to thank my parents who have always been my backbone in every step of life.

# Contents

# 1. INTRODUCTION

## 1.1 Business Problem Framing:

Airline industry is one of the most sophisticated in its use of dynamic pricing strategies to maximize revenue, based on proprietary algorithms and hidden variables. That is why the airline companies use complex algorithms to calculate the flight ticket prices. There are several different factors on which the price of the flight ticket depends. The seller has information about all the factors, but buyers are able to access limited information only which is not enough to predict the airfare prices. Considering the features such as departure time, arrival time and time of the day it will give the best time to buy the ticket. 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 models 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. With the Covid-19 impact in the market, we have seen lot of changes in the carmarket. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders,who sell used cars. With the change in market due to Covid-19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

## 1.2 Conceptual Background of the Domain Problem

Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, and it will be a different story. We might have often heard travelers saying that flight ticket prices are so unpredictable. 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 2 expensive over time

## 1.3 Review of Literature

Literature review covers relevant literature with the aim of gaining insight into the factors that are important to predict the flight ticket prices in the market. In this study, we discuss various applications and methods which inspired us to build our supervised ML techniques to predict the price of flight tickets in different locations. We did a background survey regarding the basic ideas of our project and used those ideas for the collection of data information by doing web scraping from www.yatra.comwebsite which is a web platform where buyers can book their flight tickets. This project is more about data exploration, feature engineering and preprocessing that can be done on this data. Since we scrape huge amount of data that includes more flight related features, we can do better data exploration and derive some interesting features using the available columns. Different techniques like ensemble techniques, and decision trees have been used to make the predictions. The goal of this project is to build an application which can predict the price of flight tickets with the help of other features. In the long term, this would allow people to better explain and reviewing their purchase in this increasing digital world.

## 1.4 Motivation for the Problem Undertaken

Air travel is the fastest method of transport around, and can cut hours or days off of a trip. But we know how unexpectedly the prices vary. So, I was 3 interested in Flight Fares Prediction listings to help individuals and find the right fares based on their needs. And also, to get hands on experience and to know that how the data scientist approaches and work in an industry end to end.

# 2. Analytical Problem Framing

## 2.1 Mathematical/ Analytical Modeling of the Problem

We need to develop an efficient and effective Machine Learning model which predicts the price of flight tickets. So, “Price” is our target variable which is continuous in nature. Clearly it is a Regression problem where we need to use regression algorithms to predict the results. This project is done on three phases:

• Data Collection Phase: I have done web scraping to collect the data of flights from the well-known website www.yatra.com where I found more features of flights compared to other websites and I fetch data for different locations. As per the requirement we need to build the model to predict the prices of flight tickets.

• Data Analysis: After cleaning the data I have done some analysis on the data by using different types of visualizations.

• Model Building: After collecting the data, I built a machine learning model. Before model building, have done all data pre-processing steps. The complete life cycle of data science that I have used in this project are as follows:

• Data Cleaning
• Exploratory Data Analysis
• Data Pre-processing
• Model Building
• Model Evaluation
• Selecting the best model

## 2.2 Data Sources and their formats

We have collected the dataset from the website www.yatra.com which is a web platform where the people can purchase/book their flight tickets. The data is scraped using Web scraping technique and the framework used is Selenium. We scrapped nearly 5303 of the data and fetched the data for 4 different locations and collected the information of different features of the flights and saved the collected data in excel format. The dimension of the dataset is 5303 rows and 9 columns including target variable “Price”. The particular dataset contains both categorical and numerical data type. The data description is as follows:

- Airline: The Name of airline
- Departure_time: The time when the journey starts from the source
- Time_of_arrival: Time of arrival at the destination
- Duration: Total duration taken by the flight to reach the destination from the source

- Source: The source from which the service begins
- Destination: The destination where the service ends
- Meal_availability: Availability of meals in the flight
- Number_of_stops: Total stops between the source and destination
- Price: The price of the flight ticket

## 2.3 Data Pre-processing Done

- ✓ As a first step I have scrapped the required data using selenium from yatra.com 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 null values in the dataset and I replaced them using imputation technique.
- ✓ 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.

## 2.4 Data Inputs- Logic- Output Relationships

- ✓ Since I had numerical columns, I have plotted dist plot to see the distribution of skewness in each column data.
- ✓ I have used bar plot for each pair of categorical features that shows therelation between label and independent features.
- ✓ I have used reg plot and strip plot to see the relation between numericalcolumns with target column.
- ✓ I can notice there is a linear relationship between maximum columnsand target.

## 2.5 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 — 4 GB and above
3. SSD — 250GB and above

**Software/s required**: -

1.Anaconda

**Libraries required:-**

- **import pandas as pd**: **pandas** are a popular Python-based data analysis toolkit which can be imported using `import pandas as pd`. It presents a diverse range of utilities, ranging from parsing multiple file formats to converting an entire data table into a numpy matrix array. This makes pandas a trusted ally in data science and machine learning.
- **import numpy as np**: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

- **import seaborn as sns:** Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas' data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.
- **Import matplotlib.pyplot as plt:** matplotlib. pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
  - ✓ from sklearn.preprocessing import LabelEncoder
  - ✓ from sklearn.preprocessing import StandardScaler
  - ✓ from sklearn.linear_model import LinearRegression
  - ✓ from sklearn.ensemble import RandomForestRegressor
  - ✓ from sklearn.tree import DecisionTreeRegressor
  - ✓ from xgboost import XGBRegressor
  - ✓ from sklearn.ensemble import GradientBoostingRegressor
  - ✓ from sklearn.metrics import classification_report
  - ✓ from sklearn.metrics import accuracy_score
  - ✓ from sklearn.model_selection import cross_val_score

# 3.Data Analysis and Visualization

## 3.1 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 Standardization to scale the data. After scaling we have to remove multicolinearity using VIF. Then followed by model building with all Regression algorithms

## 3.2 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 particular 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 Random Forest Regression as the 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 go through cross validation. Below is the list of Regression algorithms I have used in my project.

- Linear Regression
- Random Forest Regression
- Decision Tree Regression
- XGB Regression
- Gradient Boosting Regression

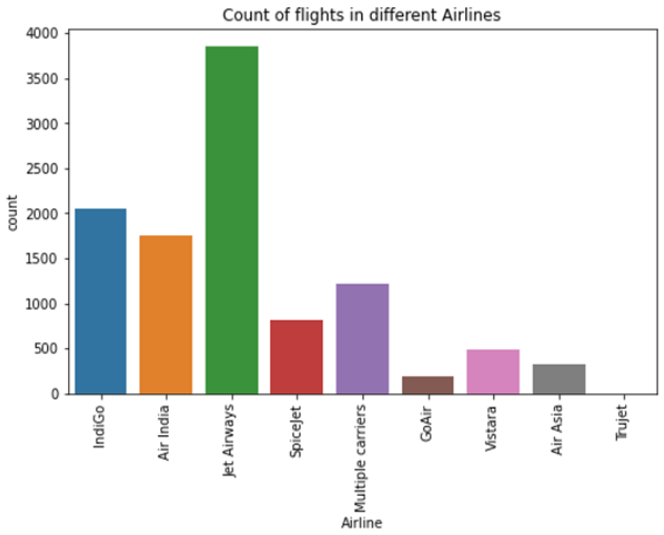
## 3.3 Key Metrics for success in solving problem under consideration

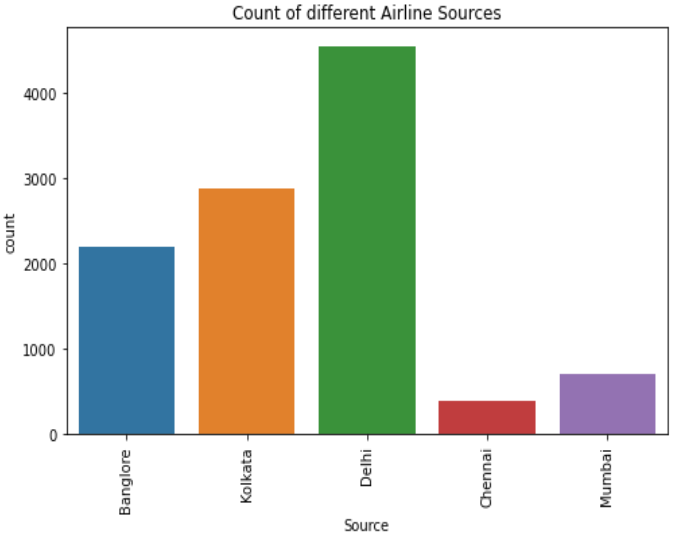
I have used the following metrics for evaluation:

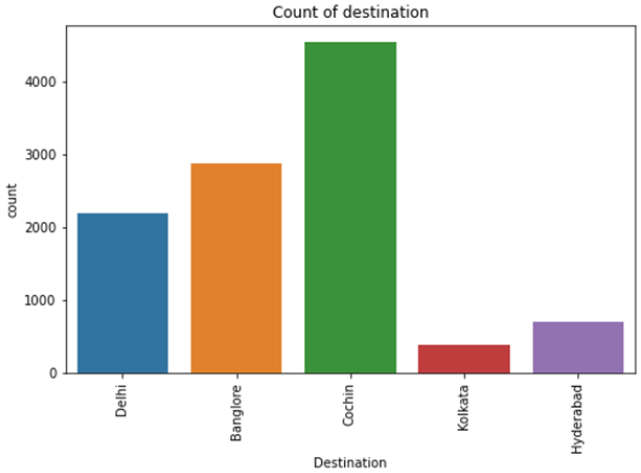
- I have used mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
- I have used root mean square deviation is one of the most commonly used measures for evaluating the quality of predictions.
- I have used r2 score which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

## 3.4 Visualizations

### 1. **Univariate Analysis:**

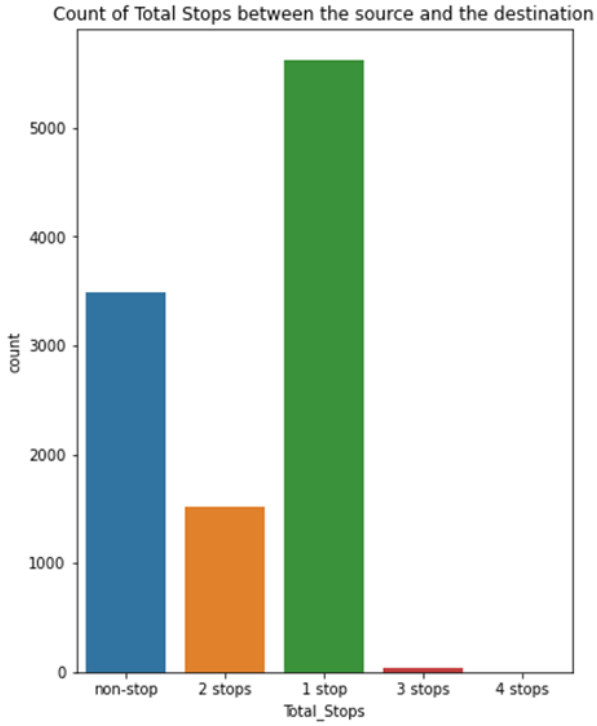


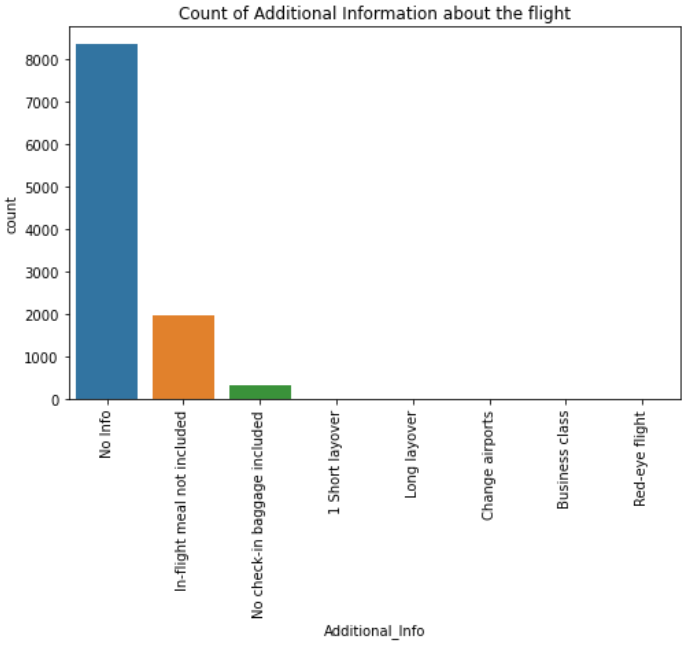



## Observations:

- From the first graph we observe that Jet Airways flights has high counts whereas Trujet and GoAir has the least counts.
- From the second graph we observe that the Journey of many flights begin from Delhi and least from Chennai
- From the third graph we observe that the destination of most of the Flights is Cochin and least is Kolkata

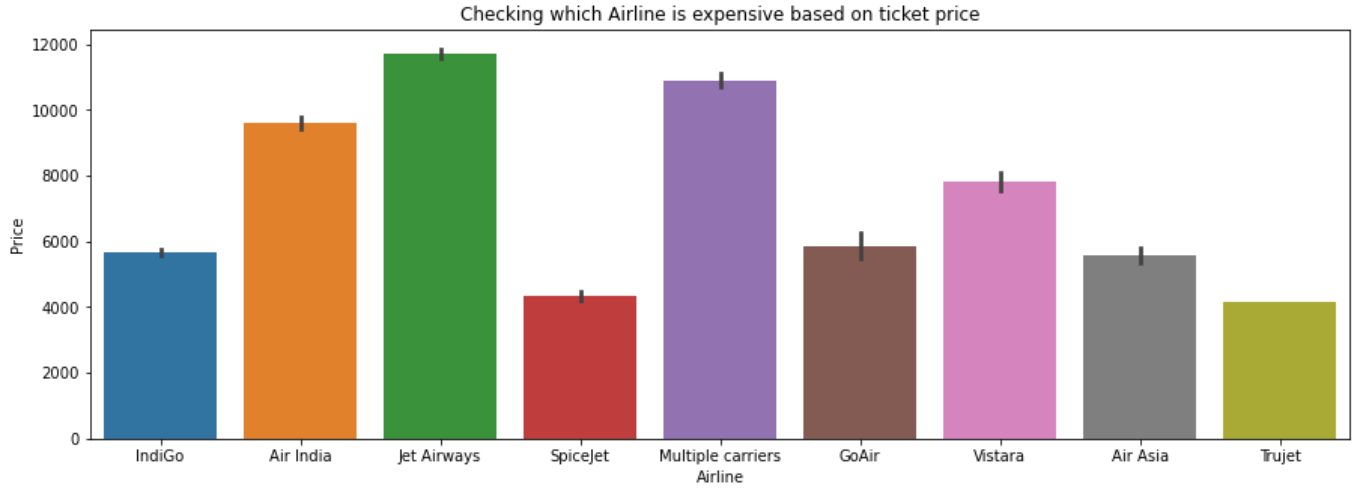


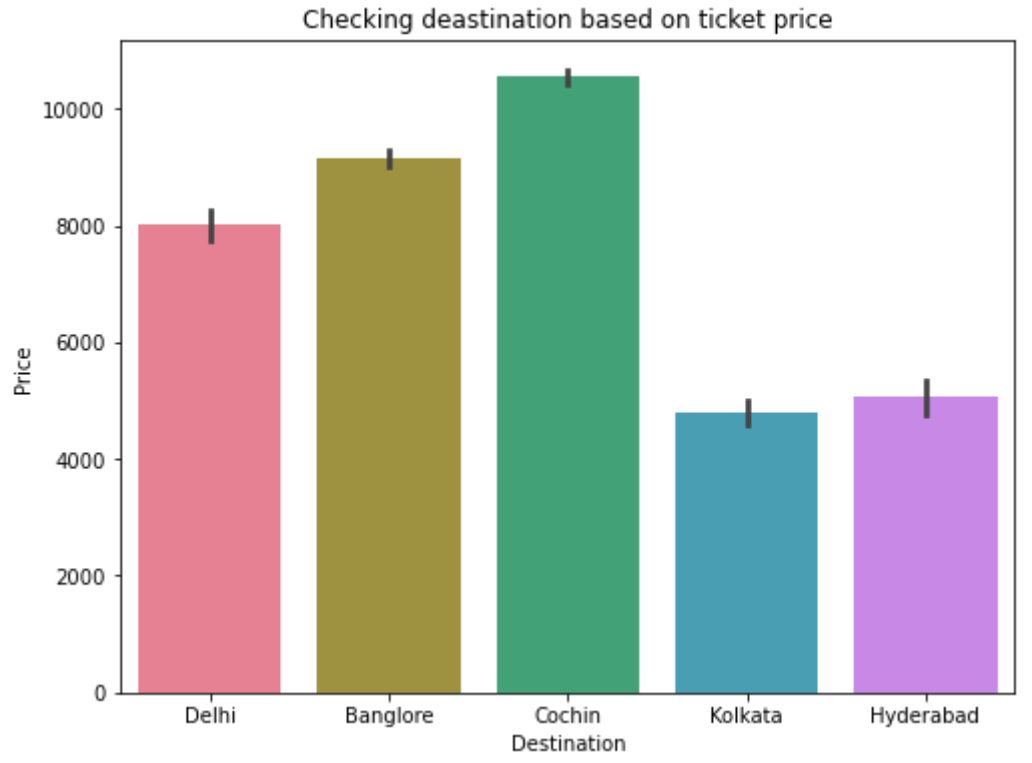




## Observations:

- From the first graph observe that most of the flights have 1 stop and very few flights have 4 stops between the source and destination.
- From the second graph we observe that most of the flights do not have any additional information

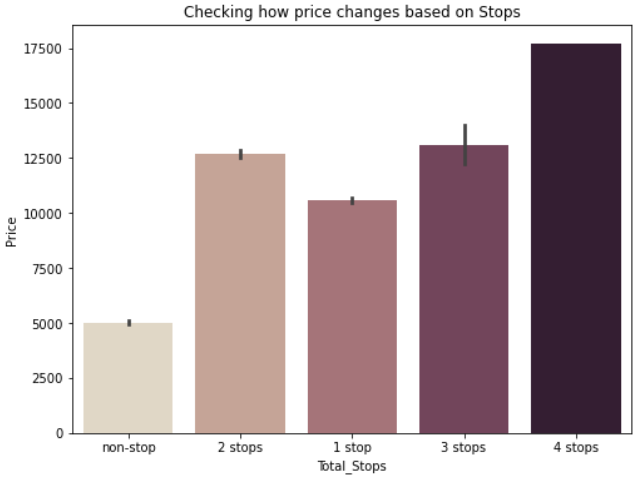
## 2. Bivariate Analysis

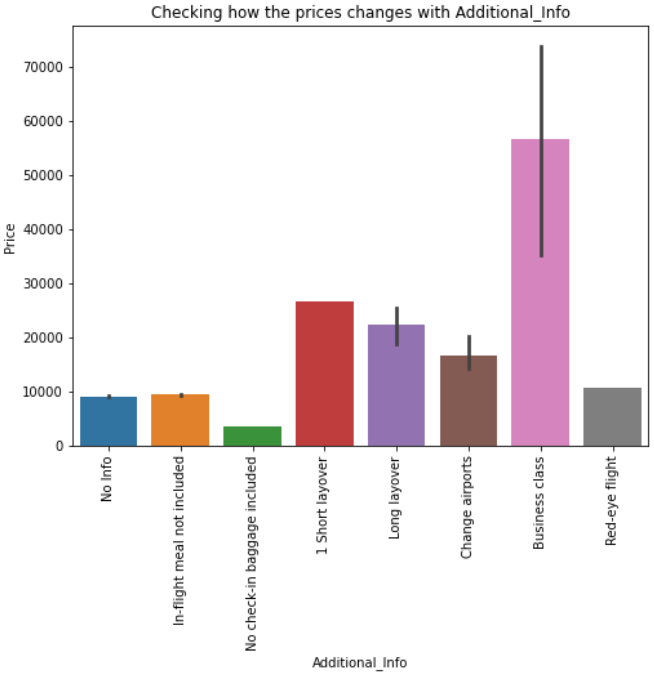


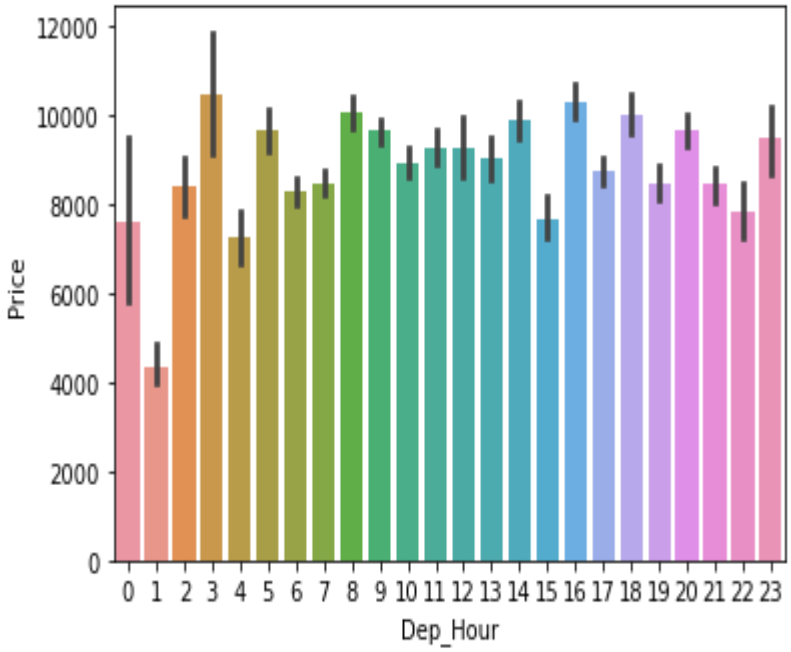




### Observations:

- The first graph from the previous slide we that Jet Airways is most expensive Airline followed by multiple carriers followed by Air India. Trujet and Spicejet have low ticket prices compared to the other Airlines.
- From the second graph on the previous slide we observe that the ticket price is the highest when the destination is Cochin

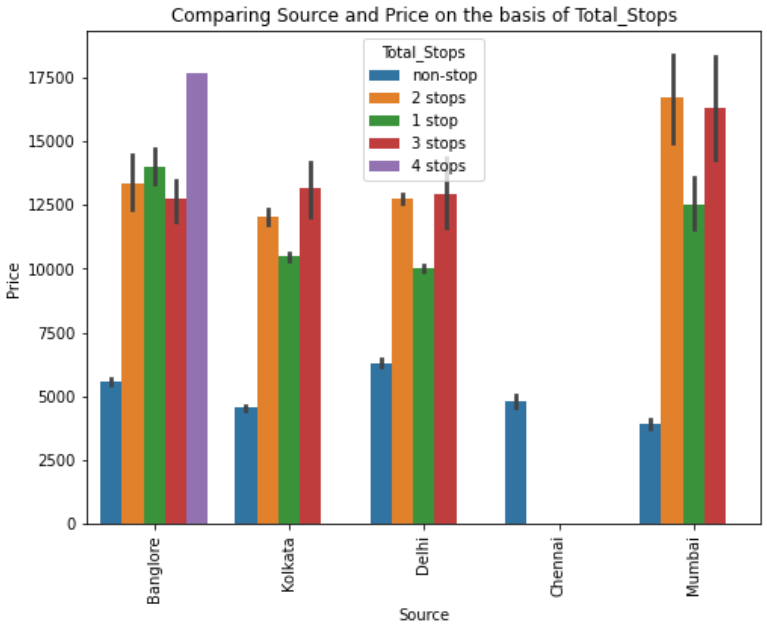



## Observations:

- From the first graph we can see that when there are 4 stops the price of the ticket is the highest and when there are no stops the price the price of the ticket is the least
- From the second graph we observe the following
  - The price of Business class tickets is the highest.
  - When there is no check-in baggage the price of the ticket is the least.
- From the third graph we can observe that Departure hour doesn't impact the ticket price much


- From the first graph we can observe the following
  - Jet Airways flights with the destination as Delhi have the highest ticket price.
  - Indigo flights with the destination as Cochin has the highest price.
  - Air India flights with the destination as Bangalore have the highest ticket price.
  - SpiceJet ticket price is the highest when the destination is Cochin
  - GoAir ticket prices are almost he same when the destination is Cochin/Banglore
  - Vistara ticket prices is the highest when the destination is Hyderabad
  - Air Asia ticket price is the highest when the destination is Cochin

- From the second graph we can observe the following
  - The flights with 4 stops and the source as Bangalore has the highest ticket price.
  - Only non-stop flights start from Chennai.
  - The price of tickets is highest when there are 2 stops are source of the flight is from Mumbai.

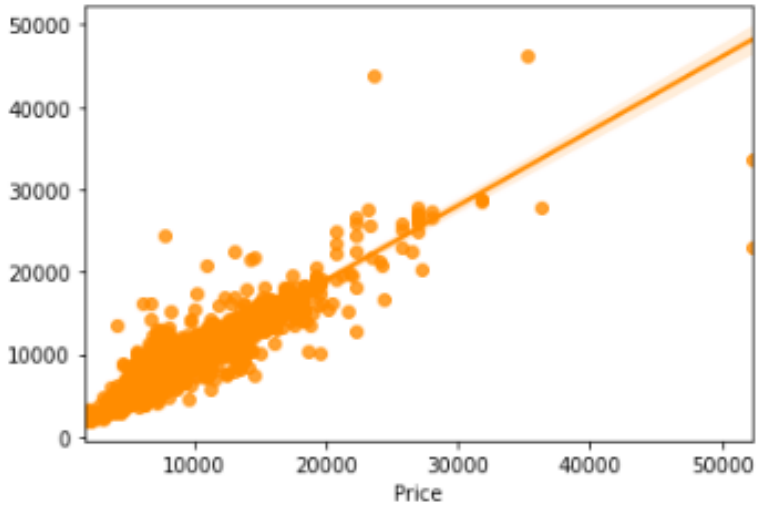
# 1. Model Building

## 1) Random Forest Regression:

```
RFR = RandomForestRegressor()
RFR.fit(x_train,y_train)
predRFR = RFR.predict(x_test)
print("R2_score:",r2_score(y_test,predRFR))
print("MAE:",metrics.mean_absolute_error(y_test,predRFR))
print("MSE:",metrics.mean_squared_error(y_test,predRFR))
print("RSME:",np.sqrt(metrics.mean_squared_error(y_test,predRFR)))
```

```
R2_score: 0.8934621988002731
MAE: 691.9220978723881
MSE: 2208444.688386039
RSME: 1486.0836747592778
```

```
sns.regplot(y_test,predRFR,color='darkorange')
plt.show()
```



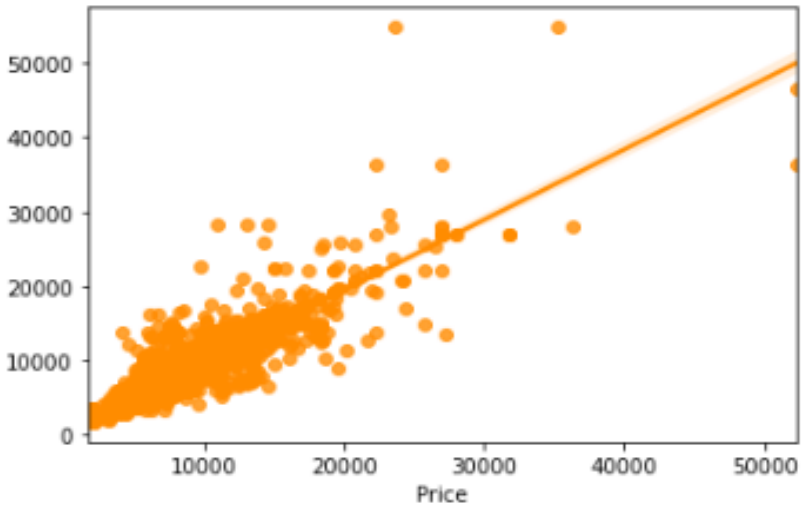

We get an r2 score of 89.34% using Random Forest Regression

## 2) Decision Tree Regression:

```
DTR = DecisionTreeRegressor()
DTR.fit(x_train,y_train)
predDTR = DTR.predict(x_test)
print("R2_score:",r2_score(y_test,predDTR))
print("MAE:",metrics.mean_absolute_error(y_test,predDTR))
print("MSE:",metrics.mean_squared_error(y_test,predDTR))
print("RSME:",np.sqrt(metrics.mean_squared_error(y_test,predDTR)))
```

```
R2_score: 0.8414954972698185
MAE: 764.7572190834903
MSE: 3285673.4717426933
RSME: 1812.6426762444642
```

```
sns.regplot(y_test,predDTR,color='darkorange')
plt.show()
```



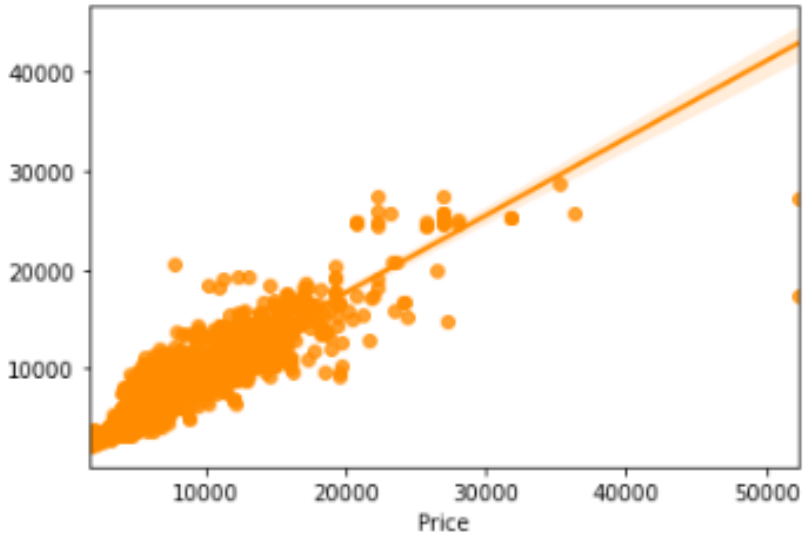

We get an r2 score of 84.14% using Decision Tree Regression

## 3) Gradient Boosting Regression:

```
GB = GradientBoostingRegressor()
GB.fit(x_train,y_train)
predGB = GB.predict(x_test)
print("R2_score:",r2_score(y_test,predGB))
print("MAE:",metrics.mean_absolute_error(y_test,predGB))
print("MSE:",metrics.mean_squared_error(y_test,predGB))
print("RSME:",np.sqrt(metrics.mean_squared_error(y_test,predGB)))
```

```
R2_score: 0.820503075372821
MAE: 1263.255162848369
MSE: 3720829.8398365923
RSME: 1928.9452661588384
```

```
sns.regplot(y_test,predGB,color='darkorange')
plt.show()
```




We get an r2 score of 82.05% using  Gradient Boosting Regression

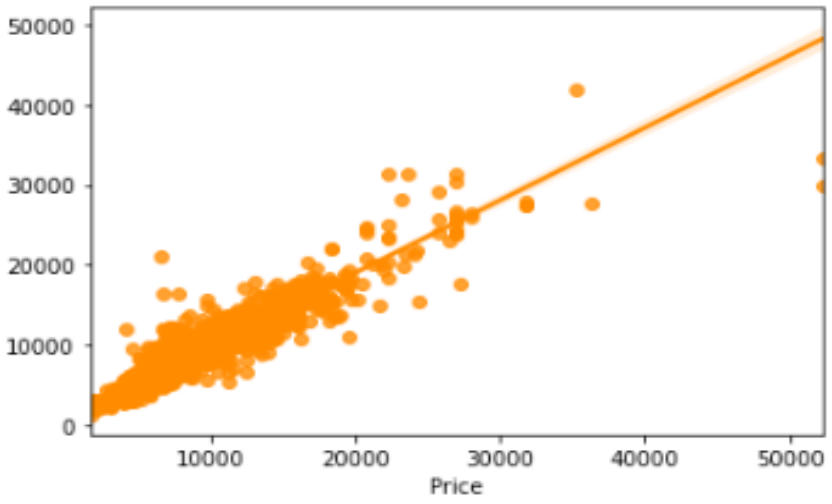
## 4) XGB Regression:

```
from xgboost import XGBRegressor as xgb

XGB=xgb(verbosity=0)
XGB.fit(x_train,y_train)
predXGB = XGB.predict(x_test)
print("R2_score:",r2_score(y_test,predXGB))
print("MAE:",metrics.mean_absolute_error(y_test,predXGB))
print("MSE:",metrics.mean_squared_error(y_test,predXGB))
print("RSME:",np.sqrt(metrics.mean_squared_error(y_test,predXGB)))
```

```
R2_score: 0.9120005978855135
MAE: 762.0725667777483
MSE: 1824158.2799005907
RSME: 1350.6140380954844
```

```
sns.regplot(y_test,predXGB,color='darkorange')
plt.show()
```




We get an r2 score of 91.2% using XGB Regression

## 2. Hyper Parameter Tuning:

```
from sklearn.model_selection import GridSearchCV
```

```
parameters = {'n_estimator':[50,100,200,400],
              'gamma':np.arange(0,0.2,0.1),
              'max_depth':[4,6,8,10],
              'n_jobs':[-2,-1,1]}
```

```
GCV = GridSearchCV(xgb(),parameters,cv=5)
```

```
GCV.fit(x_train,y_train)
```

```
GridSearchCV(cv=5,
             estimator=XGBRegressor(base_score=None, booster=None,
                                    colsample_bylevel=None,
                                    colsample_bynode=None,
                                    colsample_bytree=None,
                                    enable_categorical=False, gamma=None,
                                    gpu_id=None, importance_type=None,
                                    interaction_constraints=None,
                                    learning_rate=None, max_delta_step=None,
                                    max_depth=None, min_child_weight=None,
                                    missing=nan, monotone_constraints=None,
                                    n_estimators=100, n_jobs=None,
                                    num_parallel_tree=None, predictor=None,
                                    random_state=None, reg_alpha=None,
                                    reg_lambda=None, scale_pos_weight=None,
                                    subsample=None, tree_method=None,
                                    validate_parameters=None, verbosity=None),
             param_grid={'gamma': array([0. , 0.1]), 'max_depth': [4, 6, 8, 10],
                         'n_estimator': [50, 100, 200, 400],
                         'n_jobs': [-2, -1, 1]})
```

```
GCV.best_params_
```

```
{'gamma': 0.0, 'max_depth': 6, 'n_estimator': 50, 'n_jobs': -2}
```

```
flight_pred = xgb(gamma=0.0,max_depth=6,n_estimator=50,n_jobs=-2)
flight_pred.fit(x_train,y_train)
pred = flight_pred.predict(x_test)
print('R2_Score:',r2_score(y_test,pred)*100)
print("RMSE value:",np.sqrt(metrics.mean_squared_error(y_test, pred)))
print('MAE:',metrics.mean_absolute_error(y_test, pred))
print('MSE:',metrics.mean_squared_error(y_test, pred))
```

```
R2_Score: 91.20005978855134
RMSE value: 1350.6140380954844
MAE: 762.0725667777483
MSE: 1824158.2799005907
```

- After building the different models, and checking their cross-validation scores, I came to a conclusion that XGB Regression gave me the best R2 score
- I got R2 score of 91.2% after hyper tuning the parameters

## 5. Saving the model and Prediction:

- I have saved my best model using pickle as follows.

```
import pickle
filename = "flight_prediction.pkl"
pickle.dump(flight_pred,open(filename,'wb'))
```

# Prediction Results

```
a = np.array(y_test)
predicted = np.array(flight_pred.predict(x_test))
new_df = pd.DataFrame({"Original":a,"Predicted":predicted},index= range(len(a)))
new_df.head()
```

| | Original | Predicted |
|---|---|---|
| 0 | 17024 | 16208.969727 |
| 1 | 7817 | 9413.871094 |
| 2 | 13376 | 13726.608398 |
| 3 | 14486 | 10832.764648 |
| 4 | 11560 | 10564.657227 |

## 3.6 Interpretation of the Results

**Visualizations:** In univariate analysis I have used count plots and pie plots to visualize the counts in categorical variables and distribution plot to visualize the numerical variables. In bivariate analysis I have used bar plots, strip plots, line plots, reg plots, box plots, to check the relation between label and the features. Used pair plot to check the pairwise relation between the features. The heat map and bar plot helped me to understand the correlation between dependent and independent features. Detected outliers and skewness with the help of box plots and distribution plots respectively. And I found some of the features skewed to right as well as to left. I got to know the count of each column using bar plots.

**Pre-processing:** The dataset should be cleaned and scaled to build the ML models to get good predictions. I have performed few processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

**Model building:** After cleaning and processing data, I performed train test split to build the model. I have built multiple regression models to get the accurate R2 score, and evaluation metrics like MAE, MSE and RMSE. XGB Regression was the best model which gives 91.2% R2score. After tuning the best model, the R2 score of XGB Regression did not change. Finally, I saved my final model and got the predictions results for the price of flight tickets.

# 4. CONCLUSION

## 4.1 Key Findings and Conclusions of the Study

In this study, we have used multiple machine learning models to predict the flight ticket price. We have gone through the data analysis by performing feature engineering, finding the relation between features and label through visualizations. And got the important feature and we used these features to predict the car price by building ML models. Performed hyper parameter tuning on the best model and the best model's R2 score did not increase. We have also got good prediction results of ticket price

## 4.2 Learning Outcomes of the Study in respect of Data Science

- ✓ While working on this project I learned many things about the features of flights and about the flight ticket selling web platforms and got the idea that how the machine learning models have helped to predict the price of flight tickets.
- ✓ I found that the project was quite interesting as the dataset contains several types of data. I used several types of plotting to visualize the relation between target and features. This graphical representation helped me to understand which features are important and how these features describe price of tickets.
- ✓ Data cleaning was one of the important and crucial things in this project where I dealt with features having string values, features extraction and selection.
- ✓ Finally got XGB Regression as best model.
- ✓ The challenges I faced while working on this project was when I was scrapping the real time data from Yatra website, it took so much time to gather data.
- ✓ Finally, our aim was achieved by predicting the flight ticket price and built flight price evaluation model that could help the buyers to understand the future flight ticket prices

## 4.3 Limitations of this work and Scope for Future Work

- ✓ The main limitation of this study is the low number of records that have been used.
- ✓ In the dataset our data is not properly distributed in some of the columns many of the values in the columns are having string values which I had taken care.
- ✓ Due to some reasons our models may not make the right patterns and the performance of the model also reduces. So that issues need to be taken care.
- ✓ Finally, it would be interesting to compare our system's accuracy against that of the commercial systems available today (preferably over a period of time).