



NAME OF THE PROJECT

**FLIGHT PRICE PREDICTION USING ML
TECHNIQUES**

Submitted by:

SNEHA SANTRA

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 Ms. Khushboo Garg (SME Flip Robo), she is the person who has helped me to get out of all the difficulties I faced while doing the project.

A huge thanks to “Data trained” who are the reason behind my Internship at Fliprobo.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towardsdatascience, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)
4. Data Science Projects with Python Second Edition by Packt
5. Hands on Machine learning with scikit learn and tensor flow by Aurelien Geron

INTRODUCTION

- Business Problem Framing

The Airline Companies is considered as one of the most enlightened industries using complex methods and complex strategies to allocate airline prices in a dynamic fashion. These industries are trying to keep their all-inclusive revenue as high as possible and boost their profit. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible and maximize their profit. However, mismatches between available seats and passenger demand usually leads to either the customer paying more or the airlines company losing revenue. Airlines companies are generally equipped with advanced tools and capabilities that enable them to control the pricing process. However, customers are also becoming more strategic with the development of various online tools to compare prices across various airline companies. In addition, competition between airlines makes the task of determining optimal pricing is hard for everyone.

Time of purchase patterns (making sure last-minute purchases are expensive)

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)

So, this project involves collection of data for flight fares with other features and building a model to predict fares of flights.

- The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and the opportunity to deploy my skillset in solving a real time problem has been the primary motivation.
- Early prediction of the demand along a given route could help an airline company pre-plan the flights and determine appropriate pricing for the route. In addition, competition between airlines makes the task of determining optimal pricing is hard for everyone. So prime motive is to
- build flight price predication system based on short range timeframe (7-
 - Conceptual Background of the Domain Problem
 - The project is sub-divided following section. These are:
 - 1. Loading necessary libraries
 - 2. Loading Dataset from a CSV file
 - 3. Summarization of Data to understand Dataset (Descriptive Statistics)
 - 4. Visualization of Data to understand Dataset (Plots, Graphs etc.)
 - 5.Processing the data for modeling
 - 6.skewness and outliers detection for better accuracy
 - 7.Build the model and select the right model and save it
 - Review of Literature

On the airlines side, the main goal is increasing revenue and maximizing profit. According to (Narangajavana et al., 2014) [9], airlines utilize various kinds of pricing strategies to determine optimal ticket prices: long-term pricing policies, yield pricing which describes the impact of production conditions on ticket prices, and dynamic pricing which is mainly associated with dynamic adjustment of ticket prices in response to various influencing factors.

- Motivation for the Problem Undertaken

The project was the first provided to me by Flip Robo Technologies as a part of the internship programme. The exposure to real world data and

the opportunity to deploy my skillset in solving a real time problem has been the primary motivation.

Early prediction of the demand along a given route could help an airline company pre-plan the flights and determine appropriate pricing for the route. In addition, competition between airlines makes the task of determining optimal pricing is hard for everyone. So prime motive is to build flight price predication system based on short range timeframe (7-14 days) data available prior to actual take-off date.

Analytical Problem Framing

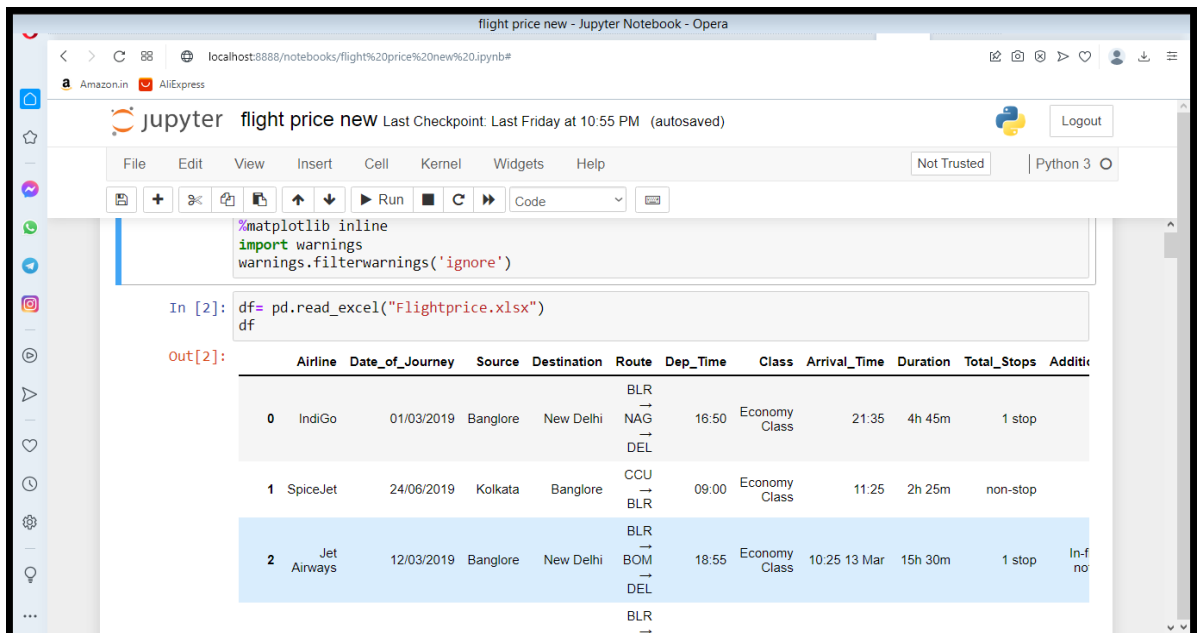
- **Mathematical/ Analytical Modeling of the Problem**

First phase of problem modelling involves data scraping of flights from internet. For that purpose, flight data is scrap from yatra.com,skyscanner.com .Data is scrape for flights for many route.Data is scrap for Economy class, Premium Economy class & Business class flights. Next phase is data cleaning & pre-processing for building ML Model. Our objective is to predict flight prices which can be resolve by use of regression-based algorithm. Further Hyperparameter tuning performed to build more accurate model out of best model.

- **Data Sources and their formats**

First phase of problem modelling involves data scraping of flights from internet. For that purpose, flight data is scrap from yatra.com,skyscanner.com .Data is scrape for flights for many

route.Data is scrap for Economy class, Premium Economy class & Business class flights.



The screenshot shows a Jupyter Notebook interface with the following code and output:

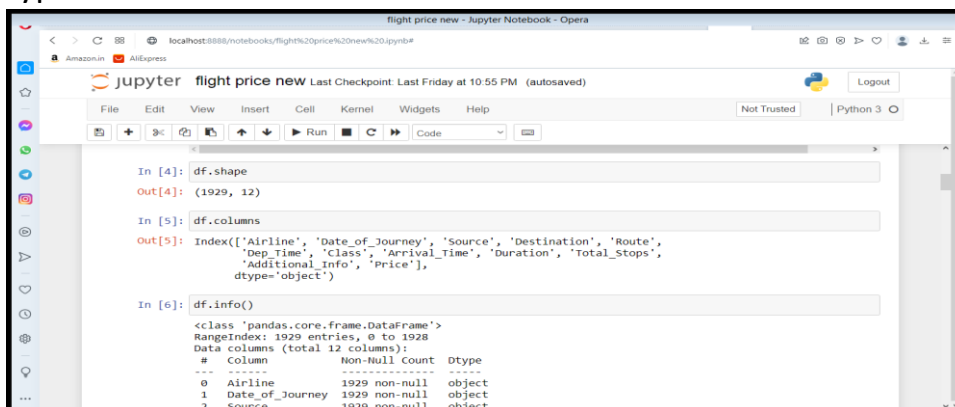
```
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')

In [2]: df = pd.read_excel("Flightprice.xlsx")
df

Out[2]:
```

| | Airline | Date_of_Journey | Source | Destination | Route | Dep_Time | Class | Arrival_Time | Duration | Total_Stops | Additional_Info |
|---|-------------|-----------------|----------|-------------|-----------------------------|----------|---------------|--------------|----------|-------------|-----------------|
| 0 | IndiGo | 01/03/2019 | Banglore | New Delhi | BLR → NAG → DEL | 16:50 | Economy Class | 21:35 | 4h 45m | 1 stop | |
| 1 | SpiceJet | 24/06/2019 | Kolkata | Banglore | CCU → BLR | 09:00 | Economy Class | 11:25 | 2h 25m | non-stop | |
| 2 | Jet Airways | 12/03/2019 | Banglore | New Delhi | BLR → BOM → DEL | 18:55 | Economy Class | 10:25 13 Mar | 15h 30m | 1 stop | In-f no |

There are 12 features in dataset including target feature 'Price'. The data types of different features are as shown below:



The screenshot shows a Jupyter Notebook interface with the following code and output:

```
In [4]: df.shape
Out[4]: (1929, 12)

In [5]: df.columns
Out[5]: Index(['Airline', 'Date_of_Journey', 'Source', 'Destination', 'Route', 'Dep_Time', 'Class', 'Arrival_Time', 'Duration', 'Total_Stops', 'Additional_Info', 'Price'], dtype='object')

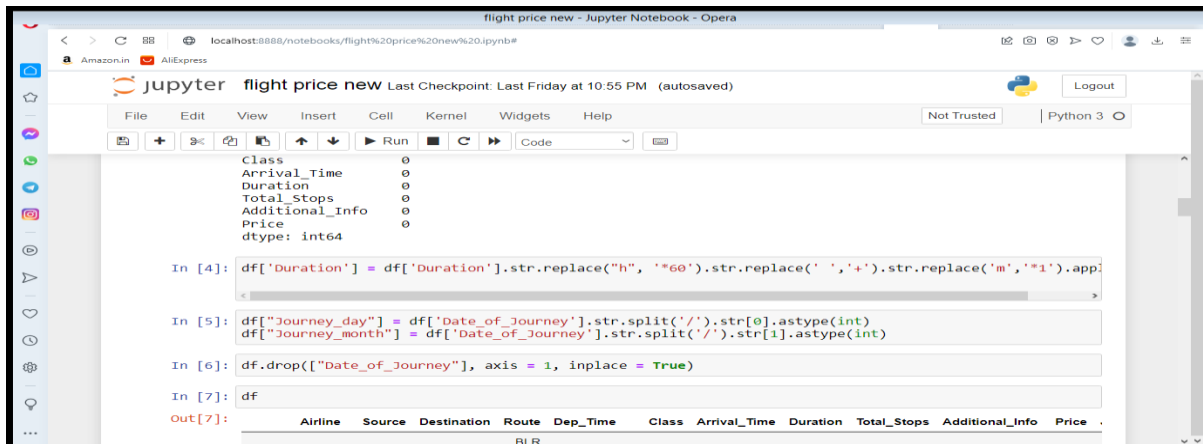
In [6]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1929 entries, 0 to 1928
Data columns (total 12 columns):
#   column              Non-Null Count  Dtype
---  ---
0   Airline              1929 non-null   object
1   Date_of_Journey      1929 non-null   object
2   Source               1929 non-null   object
```

- **Data Preprocessing Done**
The dataset is large and it may contain some data error. In order to reach clean, error free data some data cleaning & data pre-processing performed data.
- Data Integrity check –
No missing values or duplicate entries present in dataset

- Conversion of Duration column from hr & Minutes format into Minutes –

By default, Duration of flights are given in format of [(hh) hours: (mm)minute] which need to convert into uniform unit of time. Here we have written code to convert duration in terms of minute.



```

Class      0
Arrival_Time 0
Duration    0
Total_Stops 0
Additional_Info 0
Price      0
dtype: int64

In [4]: df['Duration'] = df['Duration'].str.replace("h", '*60').str.replace(':', '+').str.replace('m', '*1').astype(int)

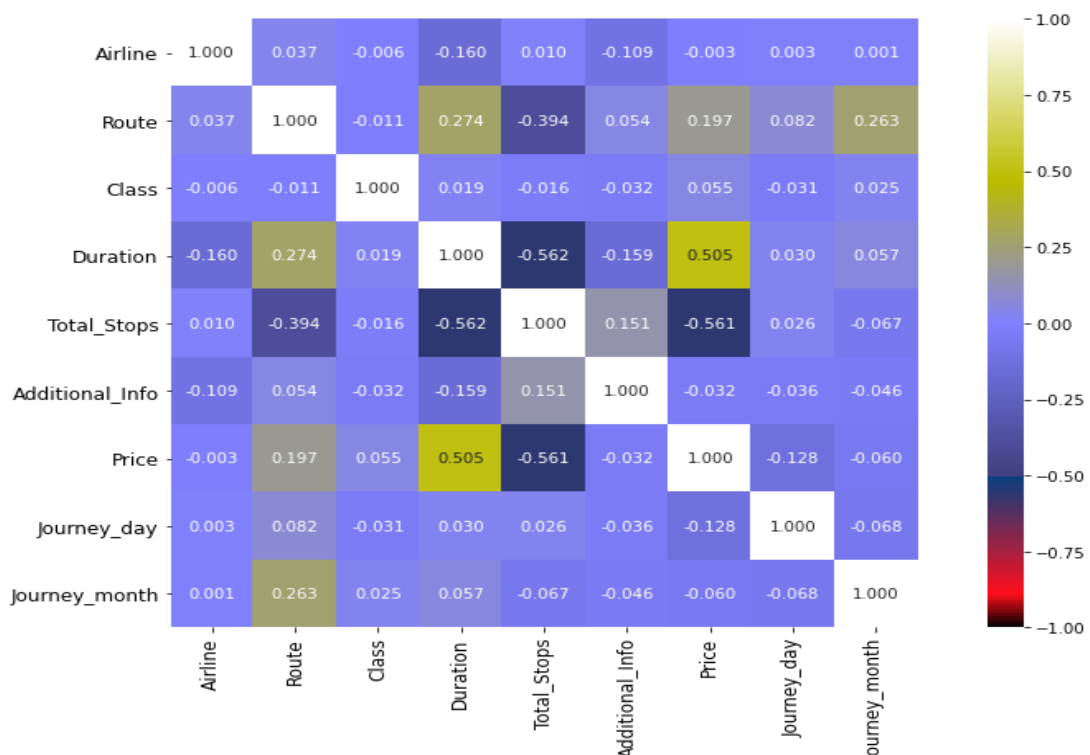
In [5]: df["Journey_day"] = df["Date_of_Journey"].str.split('/').str[0].astype(int)
df["Journey_month"] = df["Date_of_Journey"].str.split('/').str[1].astype(int)

In [6]: df.drop(["Date_of_Journey"], axis = 1, inplace = True)

In [7]: df
Out[7]:
  Airline  Source Destination  Route  Dep_Time  Class  Arrival_Time  Duration  Total_Stops  Additional_Info  Price
0  BLR

```

- Data Inputs- Logic- Output Relationships



Correlation heatmap is plotted to gain understanding of relationship between target features & independent features.

- Hardware and Software Requirements and Tools Used

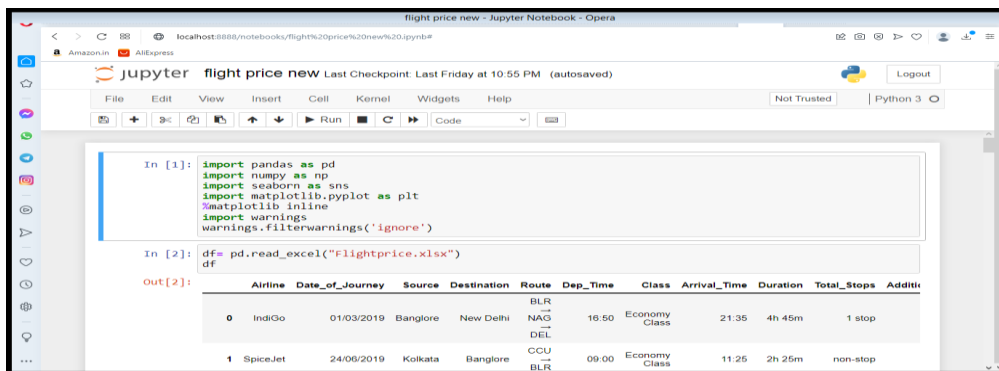
Hardware Used -

1. Processor — Intel i3 processor with 1.70GHZ
2. RAM — 12 GB

Software utilised -

1. Anaconda – Jupyter Notebook
2. Selenium – Web scraping

Libraries Used – General library for data wrangling & visualisation



The screenshot shows a Jupyter Notebook titled "flight price new" in a web browser. The code in the notebook is as follows:

```
In [1]: import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')

In [2]: df = pd.read_excel("Flightprice.xlsx")
df
```

The output of the second cell is a DataFrame with the following columns: Airline, Date_of_Journey, Source, Destination, Route, Dep_Time, Class, Arrival_Time, Duration, Total_Stops, and Additl. The first two rows of data are:

| | Airline | Date_of_Journey | Source | Destination | Route | Dep_Time | Class | Arrival_Time | Duration | Total_Stops | Additl |
|---|----------|-----------------|----------|-------------|-----------------------------|----------|------------------|--------------|----------|-------------|--------|
| 0 | IndiGo | 01/03/2019 | Banglore | New Delhi | BLR → NAG → DEL | 16:50 | Economy Class | 21:35 | 4h 45m | 1 stop | |
| 1 | SpiceJet | 24/06/2019 | Kolkata | Banglore | CCU → BLR | 09:00 | Economy Class | 11:25 | 2h 25m | non-stop | |

Libraries used for machine learning model building

```
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import ExtraTreesRegressor
from xgboost import XGBRegressor
```


Model/s Development and Evaluation

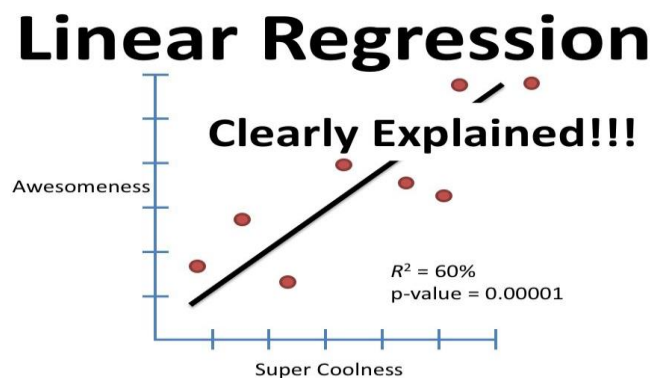
- Identification of possible problem-solving approaches (methods)

Next part of problem solving is building machine learning model to predict flight price. This problem can be solve using regression-based machine learning algorithm like linear regression. For that purpose, first task is to convert categorical variable into numerical features. Once data encoding is done then data is scaled using standard scalar. Final model is built over this scaled data. For building ML model before implementing regression algorithm, data is split in training & test data using `train_test_split` from `model_selection` module of `sklearn` library. After that model is train with various regression algorithm and 5-fold cross validation is performed. Further Hyperparameter tuning performed to build more accurate model out of best model.

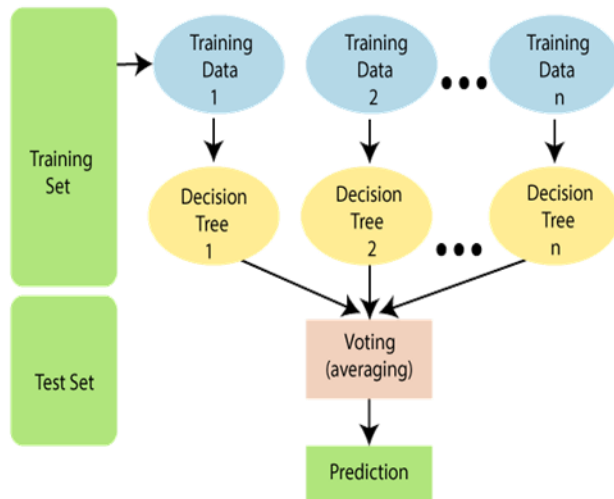
- Testing of Identified Approaches (Algorithms)

The different regression algorithm used in this project to build ML model are as below:

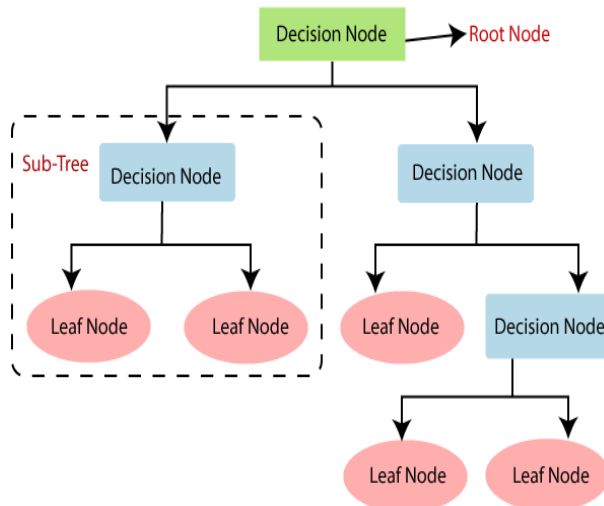
✚ Linear Regression



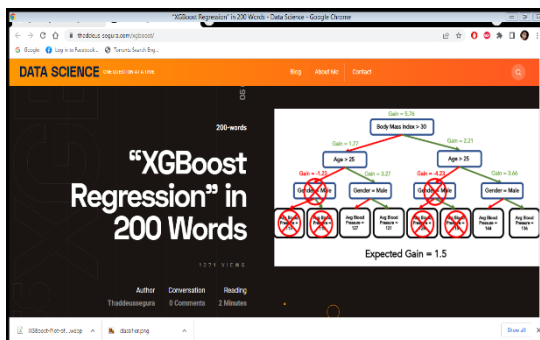
✚ Random Forest Regressor



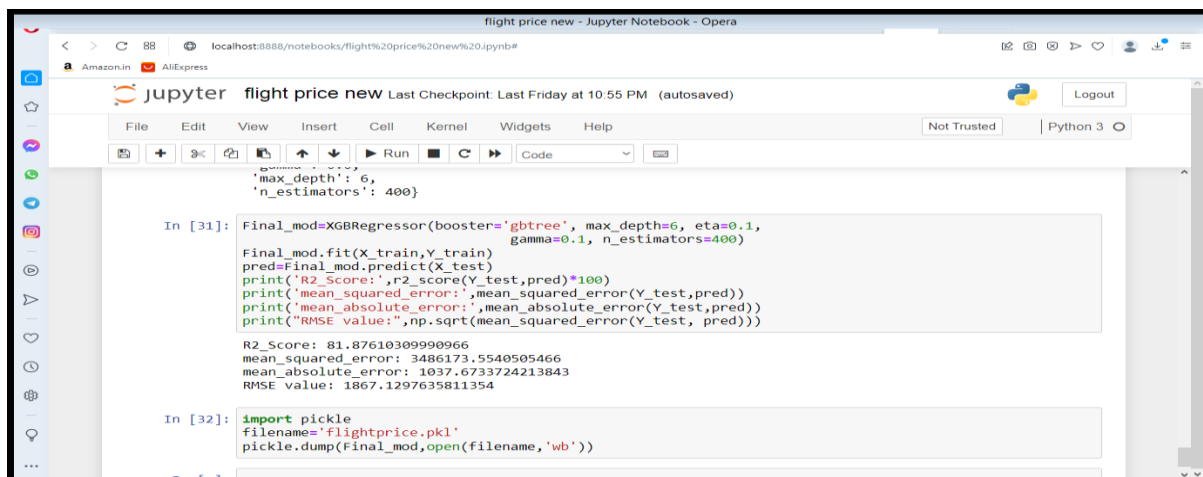
✚ Decision Tree Regressor



✚ XGB Regressor



- Run and Evaluate selected models



```
    'max_depth': 6,
    'n_estimators': 400}

In [31]: Final_mod=XGBRegressor(booster='gbtree', max_depth=6, eta=0.1,
                                gamma=0.1, n_estimators=400)
Final_mod.fit(X_train,Y_train)
pred=Final_mod.predict(X_test)
print('R2_Score:',r2_score(Y_test,pred)*100)
print('mean_squared_error:',mean_squared_error(Y_test,pred))
print('mean_absolute_error:',mean_absolute_error(Y_test,pred))
print("RMSE value:",np.sqrt(mean_squared_error(Y_test, pred)))

R2_Score: 81.87610309990966
mean_squared_error: 3486173.5540505466
mean_absolute_error: 1037.6733724213843
RMSE value: 1867.1297635811354

In [32]: import pickle
filename='flightprice.pkl'
pickle.dump(Final_mod,open(filename,'wb'))
```

5-Fold cross validation performed over all models. We can see that XGB Regressor gives maximum R2 score of 81.87 and maximum cross validation score. Among all model we will select XGB Regressor as final model and we will perform hyper parameter tuning over this model to enhance its R2 Score.

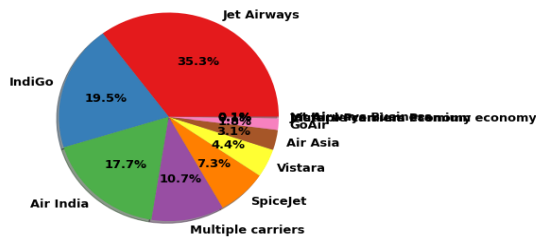
- Key Metrics for success in solving problem under consideration

Following metrics used for evaluation:

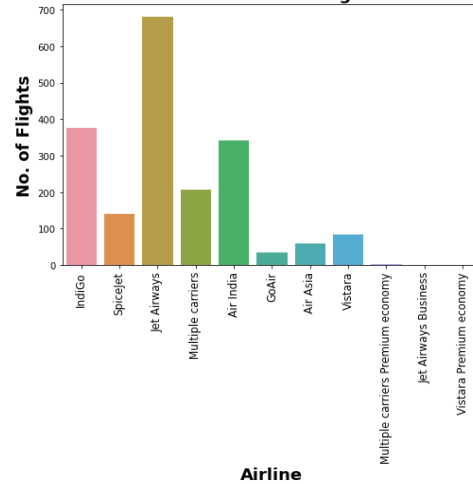
1. Mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.
2. Root mean square error is one of the most commonly used measures for evaluating the quality of predictions.
3. R2 score which tells us how accurate our model predict result, is going to important evaluation criteria along with Cross validation score.

- Visualizations

FlightWise Distribution of Airlines

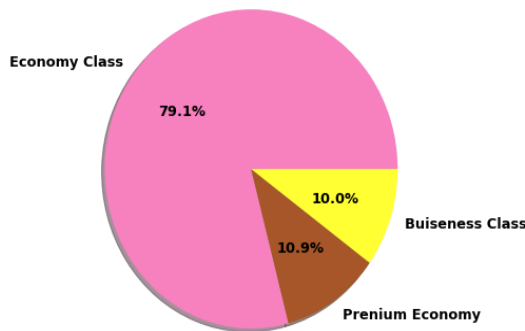


Airline Vs No of Flights

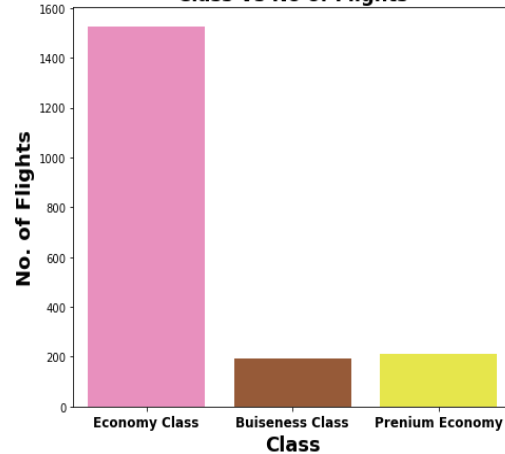


We can see maximum number of flights run by Jet Airways

Class-Wise Distribution of Flights

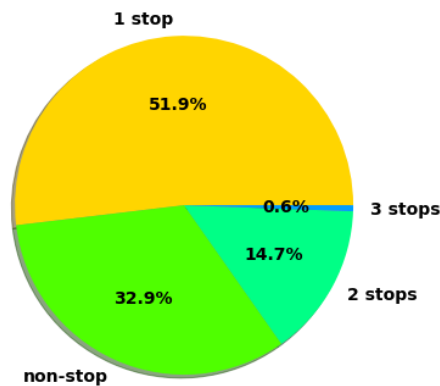


Class Vs No of Flights

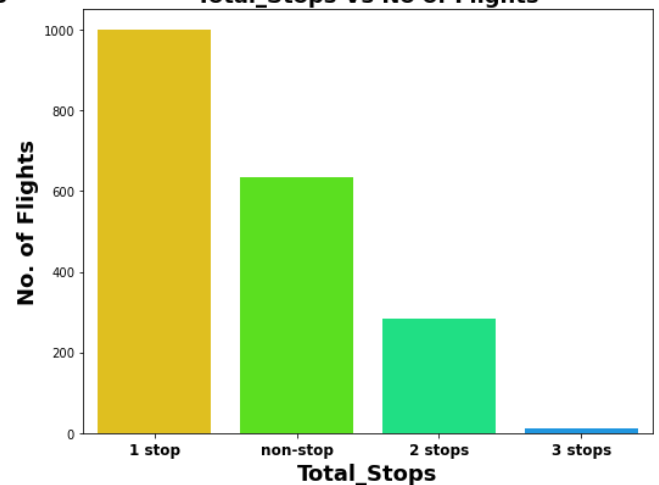


79.1% flights are of Economy class, as they are low cost of flight & most of people prefer it.

Total_Stops-Wise Distribution of Flights

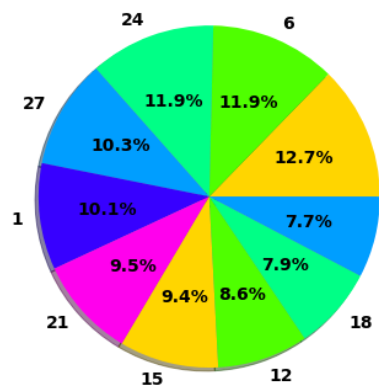


Total_Stops Vs No of Flights

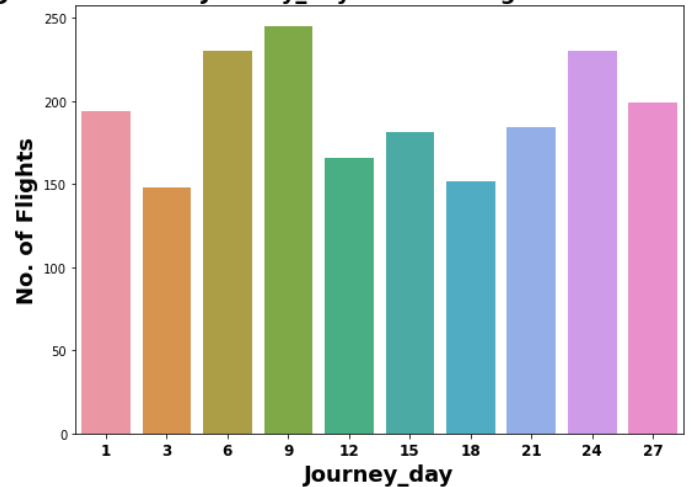


51.9% flights take single stop in there way.It is also possible that these flights may have high flight duration compare to Non-stop Flight 32.9% of flights do not have any stop in there route.

Journey_day-Wise Distribution of Flights



Journey_day Vs No of Flights



On 9th Maximum flights run while on 3rd of minimum flights run

• Interpretation of the Results

```

In [31]: Final_mod=XGBRegressor(booster='gbtree', max_depth=6, eta=0.1,
Final_mod.fit(X_train,Y_train)
pred=Final_mod.predict(X_test)
print('R2_Score:',r2_score(Y_test,pred)*100)
print('mean_squared_error:',mean_squared_error(Y_test,pred))
print('mean_absolute_error:',mean_absolute_error(Y_test,pred))
print('RMSE value:',np.sqrt(mean_squared_error(Y_test, pred)))

R2_Score: 81.87618389990966
mean_squared_error: 3486173.5540505466
mean_absolute_error: 1037.6733724213843
RMSE value: 1867.1297635811354

In [32]: import pickle
filename='flightprice.pkl'
pickle.dump(Final_mod,open(filename,'wb'))

In [ ]:

```

And we save Xgb Classifier model and accuracy is so good.

CONCLUSION

- Key Findings and Conclusions of the Study
- After the Final Submission of test data, my accuracy score was 87%
- Feature engineering helped me increase my accuracy.

- Amazingly Xgb Classifier worked better than all other Ensemble models.
- Learning Outcomes of the Study in respect of Data Science

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 car 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 prices.