NAME OF THE PROJECT

Flight Price Prediction Using ML

Submitted by:

Prathamesh Nayak

ACKNOWLEDGMENT

I would like to express my special gratitude to the “Flip Robo” team, who has given me this opportunity to deal with a beautiful dataset and it has helped me to improve my analysis 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. Last but not least my parents have been my backbone in every step of my life.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towards data science, 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
6. B. Smith, J. Leimkuhler, R. Darrow, and Samuels, “Yield management at American airlines, “Interfaces, vol. 22, pp. 8–31, 1992
7. William Groves, Maria Gini, “An agent for optimizing airline ticket purchasing”, in international conference on Autonomous agents and multi-agent systems, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2013)
8. Chen, Y., Cao, J., Feng, S., Tan, Y., 2015. An ensemble learning based approach for building airfare forecast service. In: 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 964-969.
9. Yeamduan Narangajavana, Fernando.J. Garrigos-Simon, Javier Sanchez García, Santiago Forgas-Coll, “Prices, prices and prices: A study in the airline sector”, Tourism Manage., 41 (2014), pp. 28-42

10. Bo An, Haipeng Chen, Noseong Park, V.S. Subrahmanian MAP: Frequency-Based Maximization of Airline Profits based on an Ensemble Forecasting Approach Proceedings of the 22nd ACM 3 Flight Price Prediction Using ML Techniques SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), ACM, New York, NY, USA (2016), pp. 421-430
11. R. Ren, Y. Yang, and S. Yuan, "Prediction of airline ticket price," University of Stanford, 2014.
12. T. Janssen, T. Dijkstra, S. Abbas, and A. C. van Riel, "A linear quantile mixed regression model for prediction of airline ticket prices," Radboud University, 2014.
13. K. Tziridis, T. Kalampokas, G. A. Papakostas, and K. I. Diamantaras, "Airfare prices prediction using machine learning techniques," in the 25th IEEE European signal processing conference, 2017, pp. 1036– 1039.
14. C. Koopmans and R. Lieshout, "Airline cost changes: To what extent are they passed through to the passenger?" Journal of Air Transport Management, vol. 53, pp. 1–11, 2016.
15. G. Francis, A. Fidato, and I. Humphreys, "Airport–airline interaction: the impact of low-cost carriers on two European airports," Journal of Air Transport Management, vol. 9, no. 4, pp. 267–273, 2003.
16. Boruah A., Baruah K., Das B., Das M.J., Gohain N.B. (2019) "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter," https://doi.org/10.1007/978-981-13-0224-4_18
17. G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-Based kernelized slackmin algorithm," doi:10.1016/j.jpdc.2016.09.001
18. T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.
19. Stacey Mumbower, Laurie A. Garrow, Matthew J. Higgins "Estimating flight-level price elasticities using online airline data: a first step toward integrating pricing, demand, and revenue optimization", Transportation Res. Part A: Policy Practice, 66 (2014), pp. 196-212

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 airline losing revenue. Airlines 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 difficult for everyone.

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

- **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.

Conceptual Background of the Domain Problem

A report says India's affable aeronautics industry is on a high development movement. India is the third-biggest avionics showcase in 2020 and the biggest by 2030. Indian air traffic is expected to cross 100 million travelers by 2017, whereas there were just 81 6 million passengers in 2015. Agreeing to Google, the expression" Cheap Air Tickets" is most sought in India. At the point when the white-collar class of India is presented to air travel, buyers are searching at modest costs.

Any individual who has booked a flight ticket previously knows how dynamically costs change. Aircraft uses advanced strategies called Revenue Management to execute a distinctive valuing strategy [6]. The least expensive accessible ticket changes over a period the cost of a ticket might be high or low. This valuing method naturally modifies the toll as per the time like morning, afternoon or night. Cost may likewise change with the seasons like winter, summer and celebration seasons. The extreme goal of the carrier is to build its income yet on the opposite side the purchaser is searching at the least expensive cost. Purchasers generally endeavor to purchase the ticket in advance to the take-off day.

From the customer point of view, determining the minimum price or the best time to buy a ticket is the key issue. The concept of "tickets bought in advance are cheaper" is no longer working (William Groves and Maria Gini, 2013) [7]. It is possible that customers who bought a ticket earlier pay more than those who bought the same ticket later. Moreover early purchasing implies a risk of commitment to a specific schedule that may need to be changed usually for a fee. Most of the studies performed on the customer side focus on the problem of predicting optimal ticket purchase time using statistical methods. As noted by Y. Chen et al. (2015) [8], predicting the actual ticket price is a more difficult task than predicting an optimal ticket purchase time due

to various reasons: absence of enough datasets, external factors influencing ticket prices, dynamic behavior of ticket pricing, competition among airlines, proprietary nature of airlines ticket pricing policies etc.

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. Existing demand prediction models generally try to predict passenger demand for a single flight/route and market share of an individual airline. Price discrimination allows an airline company to categorize customers based on their willingness to pay and thus charge them different prices. Customers could be categorized into different groups based on various criteria such as business vs leisure, tourist vs normal traveler, profession etc. For example, business customers are willing to pay more as compared to leisure customers as they rather focus on service quality than price.
In a less competitive market, the market power of a given airline is stronger, and thus, it is more likely to engage in price discrimination. On the other hand, the higher the level of competition, the weaker the market power of an airline, and then the less likely the chance of the airline fare increases.

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.

Among the recent work performed on route demand and market share prediction is the study done by (Bo An et al., 2016) [10]. The authors

proposed a data mining technique designed for Maximizing Airline Profits (MAP) through prediction of total route demand and market share of an individual airline. Unlike most other works, this work considers a broad set of routes (around 700 routes) across 13 airlines operating in those routes. The training dataset spans 10 years (40 quarters) while the testing set includes the first quarter of 2015 (a total of 9100 predictions). However, the prediction is performed quarterly and not for a short period of time which might not consider dynamic demand changes. Moreover, the routes considered are only national routes in the US.

Ren et al. [11] proposed using LR, Naive Bayes, Soft-max regression, and SVMs to build a prediction model and classify the ticket price into five bins (60% to 80%, 80% to 100%, 100% to 120%, and etc.) to compare the relative values with the overall average price. More than nine thousand data points, including six features (e.g., the departure week begin, price quote date, the number of stops in the itinerary, etc.), were used to build the models. The authors reported the best training error rate close to 22.9% using the LR model. Their SVM regression model failed to produce a satisfying result. Instead, an SVM classification model was used to classify the prices into either "higher" or "lower" than the average.

n [12], four LR models were compared to obtain the best fit model, which aims to provide an unbiased information to the passenger whether to buy the ticket or wait longer for a better price. The authors suggested using linear quantile mixed models to predict the lowest ticket prices, which are called the "real bargains". However, this work is limited to only one class of tickets, economy, and only on one direction single leg flights from San Francisco Airport to John F. Kennedy Airport.

Tziridis et al. [13] applied eight machine learning models, which included ANNs, RF, SVM, and LR, to predict ticket prices and compared their performance. The best regression model achieved an

accuracy of 88%. In their comparison, Bagging Regression Tree is identified as the best model, which is robust and not affected by using different input feature sets.

Macroeconomic data, such as crude oil price and Consumer Price Index (CPI), can also be utilized to uncover the hidden trend in airline fares. Fuel costs can take up to 50% of the total operating cost of an airline [14]. Hence, the level of crude oil price plays an essential rule of formulating the airline's pricing strategy. It is a common practice for airlines to pass the cost of aviation fuel to the customer by adjusting the fare to compensate for the fluctuation of crude oil price.

The emergence of Low-Cost Carrier (LCC) has revolutionized the entire operating model of the airline industry. The presence of LCC in a market has had a substantial impact on the total passenger volume and the air ticket price [15].

In detail monitoring, the passenger gets an approximation of plane price with date to choose the best blend of date and price. The price for weekend on Sunday is not possible to calculate in this presented model, as weekend on Sundays the most accidental price difference compared to other days in the week and needs more elements, a nonlinear model for successful forecast which will be the upcoming range of study to be done for this presented technique [16]. To forecast the mean plane ticket amount in the business area, machine learning support was evolved. Selecting feature techniques authors have presented a model to forecast the mean flight amount with an R squared score of 80% accuracy.

The accuracy of the logistic regression model is up to 70-75%. The conclusion of the given model is that most of the plane ticket prices vary from day to day. Authors have reported that the ticket price is high for a certain period and then it gradually decreases to a certain level. When the flight is at a difference of 2-3 days' time the ticket price starts increasing again [17].

Janssen [18] built up an expectation model utilizing the Linear Quantile Blended Regression strategy for San Francisco to New York course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the take-off date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the take-off date, anyway for a considerable length of time close to the take-off date, the expectation isn't compelling.

Business class flights are more inelastic as compared to leisure class as business customers have less flexibility to change or cancel their travel date (Mumbower et al., 2014) [19]. In contrast, short distance flights are more elastic (more price sensitive) than long distance flights because of the availability of other travel options (e.g., bus, train, car etc.). Airlines use price elasticity information to determine when to increase ticket prices or when to launch promotions so that the overall demand is increased

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 skills 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 difficult for everyone. So the prime motive is to build a flight price prediction 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 modeling involves data scraping of flights from the internet. For that purpose, flight data is scrapped from www.yatra.com for the time frame of 23 Jan 2022 to 4 Feb 2022. Data is scrape for flights on route of New Delhi to Mumbai. 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 a more accurate model out of the best model.

Data Sources and their formats

Data is collected from www.yatra.com for the time frame of 19 sep 2022 to 5 Oct 2022 using selenium and saved in a xlsx file. Data is scrape for flights on route of New Delhi to Mumbai. Data is scrap for Economy class, Premium Economy class & Business class flights. Around 3000 flight details are collected for this project.

```
# Importing dataset excel file using pandas.
df= pd.read_excel('Flight_Price_dataset.xlsx')
```

```
print('No. of Rows :',df.shape[0])
print('No. of Columns :',df.shape[1])
pd.set_option('display.max_columns',None) # This will enable us to see truncated columns
df.head()
```

```
No. of Rows : 3073
No. of Columns : 10
```

There are 10 features in the dataset including target feature 'Price'. The data types of different features are as shown below

```
# lets sort columns by their datatype
df.columns.to_series().groupby(df.dtypes).groups
```

```
{int64: ['Price'], object: ['Airline', 'Aeroplane', 'Date', 'Departure_Time', 'Arrival_Time', 'Source', 'Destination', 'Stops',
'Duration']}
```

Data Pre-processing

The dataset is large and it may contain some data errors. 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 a uniform unit of time. Here we have written code to convert duration in terms of minute. For example,

```
# Conversion of Duration column from hr & Minutes format to Minutes
df['Duration'] = df['Duration'].str.replace('h','*60').str.replace(' ','+').str.replace('m','*1').apply(eval)

# convert this column into a numeric datatypes
df['Duration']= pd.to_numeric(df['Duration'])
```

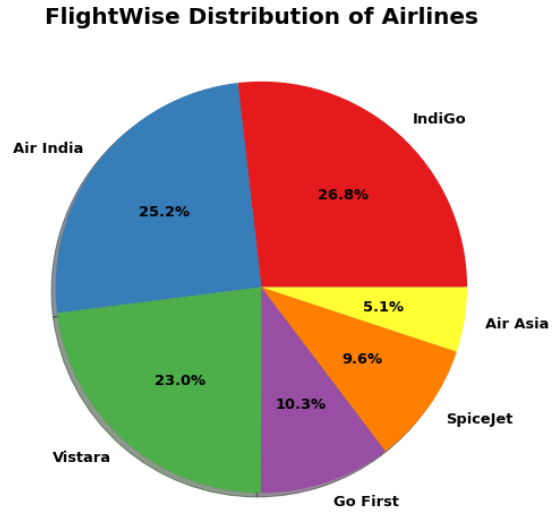
• Create a new column for day & date – New column for 'Day' & 'Date' is extracted from Date column.

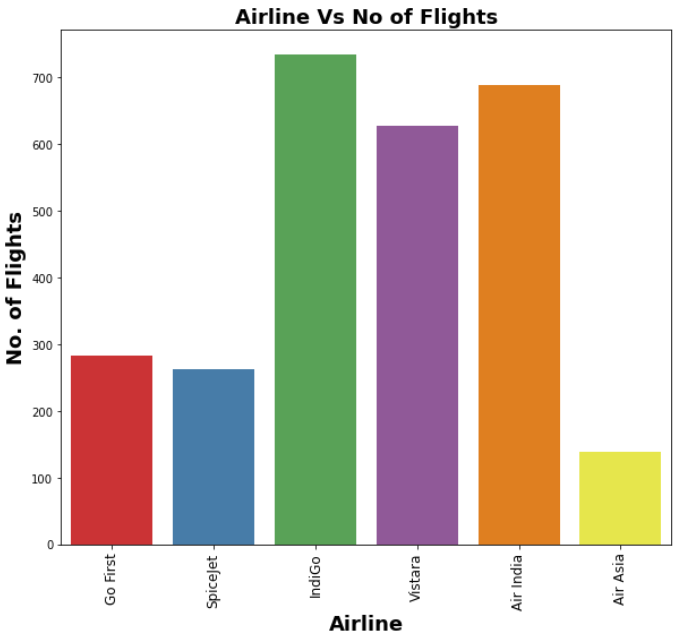
```
df['Day']= df['Date'].map(lambda x :x[:3])
```

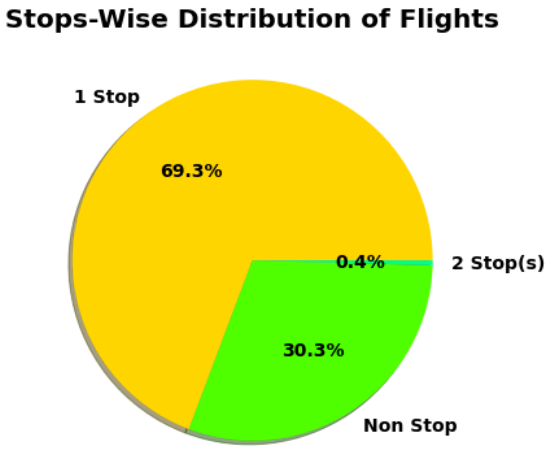
```
df['Date']= df['Date'].map(lambda x :x[4:])
```

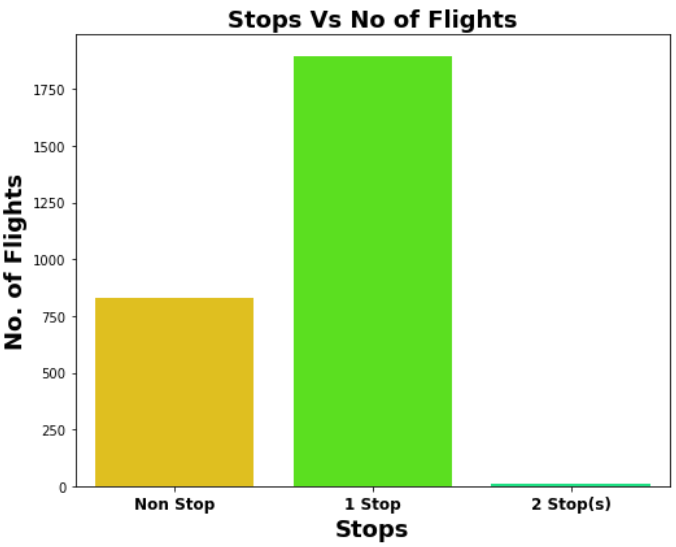
Exploratory Data Analysis

Let see key results from EDA, start with flight-wise distribution of airlines.



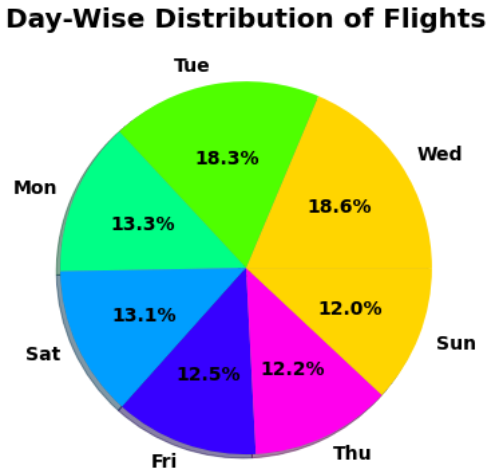


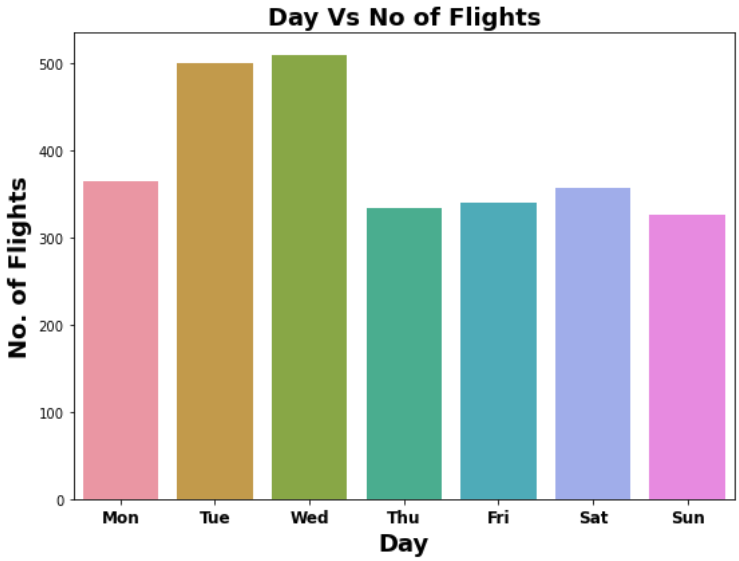




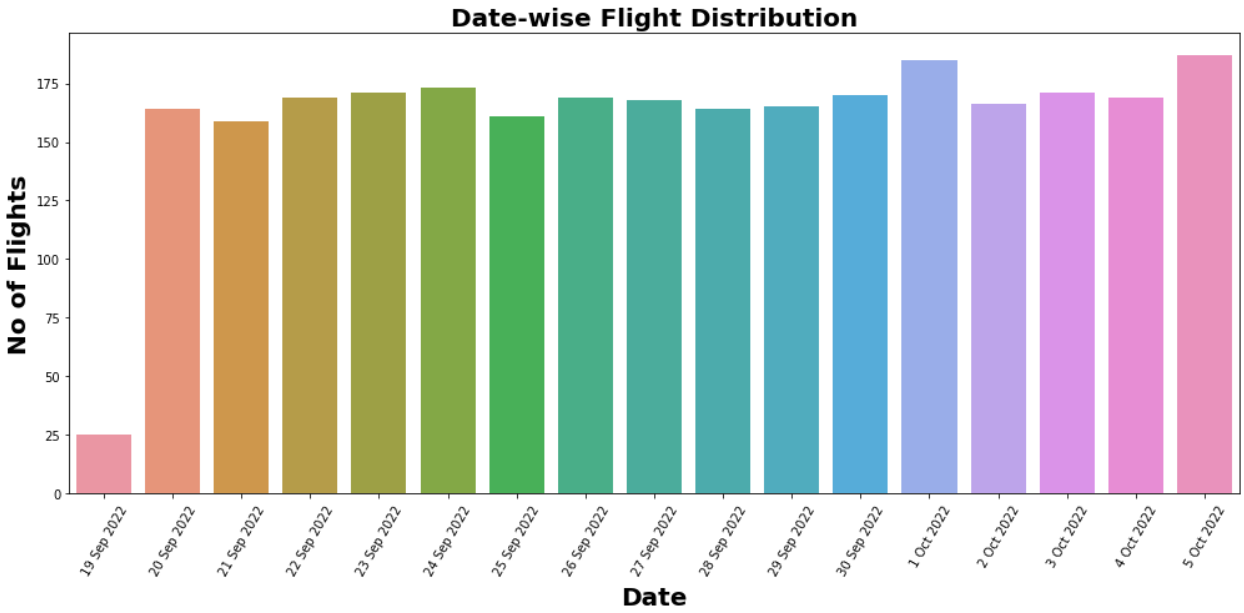
69.3% flights take single stop in there way from New Delhi to Mumbai.It is also possible that these flights may have high flight duration compare to Non-stop Flight

30.3% of flights do not have any stop in their route.





On Wednesday Maximum flights run while on Saturday minimum flights run.



We can see those Maximum flights schedule on 1 Oct 2022 & Minimum flights schedule on 19 Sep 2022

Models Development & Evaluation

❖Identification Of Possible Problem-Solving Approaches (Methods)

First part of problem solving is to scrape data from the www.yatra.com website which we have already done. Next part of problem solving is building a machine learning model to predict flight price. This problem can be solved using a regression-based machine learning algorithm like linear regression. For that purpose, the first task is to convert categorical variables 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 models before implementing a regression algorithm, data is split in training & test data using train_test_split from model_selection module of sklearn library. After that the model is trained with various regression algorithms and 5-fold cross validation is performed. Further Hyperparameter tuning performed to build a more accurate model out of the best model.

❖ Testing of Identified Approaches (Algorithms)

1. Selenium will be used for web scraping data from www.yatra.com
2. Flights on route from New Delhi to Mumbai in duration of 23 Jan 2022 to 4 Feb 2022.
3. Data is scrap in three parts:
- **Economy class flight price extraction**
- **Business class flight price extraction**
- **Premium Economy class price extraction**

4. Selecting features to be scraped from the website.
5. In the next part web scraping code executed for above mentioned details.
6. Exporting final data in Excel file

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**
- ❖ **Extra Tree Regressor**

KEY METRICS FOR SUCCESS IN SOLVING PROBLEM UNDER CONSIDERATION

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 accurately our model predicts the result, is going to be an important evaluation criteria along with Cross validation score.

Linear Regression

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state= 34, test_size=0.2)
lin_reg= LinearRegression()
lin_reg.fit(X_train, Y_train)
y_pred = lin_reg.predict(X_test)
print('\033[1m'+ 'Error :'+ '\033[0m')
print('Mean absolute error :', mean_absolute_error(Y_test,y_pred))
print('Mean squared error :', mean_squared_error(Y_test, y_pred))
print('Root Mean squared error :', np.sqrt(mean_squared_error(Y_test, y_pred)))
print('\033[1m'+' R2 Score :'+'\033[0m')
print(r2_score(Y_test,y_pred)*100)
```

```
Error :
Mean absolute error : 1346.4714650133378
Mean squared error : 3657402.548861266
Root Mean squared error : 1912.433671754727
 R2 Score :
47.93261953393117
```

Random Forest Regressor

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state= 34, test_size=0.2)
rfc = RandomForestRegressor()
rfc.fit(X_train, Y_train)
y_pred = rfc.predict(X_test)
print('\033[1m'+ 'Error of Random Forest Regressor:'+ '\033[0m')
print('Mean absolute error :', mean_absolute_error(Y_test,y_pred))
print('Mean squared error :', mean_squared_error(Y_test, y_pred))
print('Root Mean squared error :', np.sqrt(mean_squared_error(Y_test, y_pred)))
print('\033[1m'+'R2 Score of Random Forest Regressor :'+'\033[0m')
print(r2_score(Y_test,y_pred)*100)
```

```
Error of Random Forest Regressor:
Mean absolute error : 765.6459306569343
Mean squared error : 1682120.0872093067
Root Mean squared error : 1296.9657232206666
R2 Score of Random Forest Regressor :
76.05306350606304
```

Decision Tree Regressor:

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state= 34, test_size=0.2)
dtc = DecisionTreeRegressor()
dtc.fit(X_train, Y_train)
y_pred = dtc.predict(X_test)
print('\033[1m'+ 'Error of Decision Tree Regressor:'+ '\033[0m')
print('Mean absolute error :', mean_absolute_error(Y_test,y_pred))
print('Mean squared error :', mean_squared_error(Y_test, y_pred))
print('Root Mean squared error :', np.sqrt(mean_squared_error(Y_test, y_pred)))
print('\033[1m'+'R2 Score of Decision Tree Regressor :'+'\033[0m')
print(r2_score(Y_test,y_pred)*100)
```

```
Error of Decision Tree Regressor:
Mean absolute error : 781.8594890510949
Mean squared error : 2621700.4945255476
Root Mean squared error : 1619.1666049315454
R2 Score of Decision Tree Regressor :
62.677043258734685
```

Extra Trees Regressor:

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state= 34, test_size=0.2)
etc = ExtraTreesRegressor()
dtc.fit(X_train, Y_train)
y_pred = dtc.predict(X_test)
print('\033[1m'+ 'Error of Extra Tree Regressor:'+ '\033[0m')
print('Mean absolute error :', mean_absolute_error(Y_test,y_pred))
print('Mean squared error :', mean_squared_error(Y_test, y_pred))
print('Root Mean squared error :', np.sqrt(mean_squared_error(Y_test, y_pred)))
print('\033[1m'+'R2 Score of Extra Tree Regressor :'+'\033[0m')
print(r2_score(Y_test,y_pred)*100)
```

```
Error of Extra Tree Regressor:
Mean absolute error : 814.6806569343066
Mean squared error : 2752364.848540146
Root Mean squared error : 1659.0252706152928
R2 Score of Extra Tree Regressor :
60.81688415867899
```

XGB Regressor:

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scale, Y, random_state= 34, test_size=0.2)
xgb = XGBRegressor()
xgb.fit(X_train, Y_train)
y_pred = xgb.predict(X_test)
print('\033[1m'+ 'Error of XGB Regressor:'+ '\033[0m')
print('Mean absolute error :', mean_absolute_error(Y_test,y_pred))
print('Mean squared error :', mean_squared_error(Y_test, y_pred))
print('Root Mean squared error :', np.sqrt(mean_squared_error(Y_test, y_pred)))
print('\033[1m'+'R2 Score of XGB Regressor :'+'\033[0m')
print(r2_score(Y_test,y_pred)*100)
```

```
Error of XGB Regressor:
Mean absolute error : 804.1185151260264
Mean squared error : 1923901.283536315
Root Mean squared error : 1387.0476861075524
R2 Score of XGB Regressor :
72.6110268774674
```

Limitations of this work and Scope for Future Work

➢ **In this study we focus on flights on route of New Delhi to Mumbai, more route can incorporate in this project to extend it beyond present investigation.**
➢ **This investigation focus on short timeframe (14 days prior flights take off) which can be extended variation over larger period.**
➢ **Time series analysis can be performed over this model.**