Predicting the Flight Ticket Price

Introduction:

Airline carriers execute dynamic valuing for their tickets, and base their estimating choices on request prediction models. The explanation behind such a complex framework is, that each flight just has a set number of seats to sell, so aircraft need to manage the request. For the situation where the request is relied upon to surpass the limit, the carrier may build costs, to diminish the rate at which seats fill. Then again, a seat that goes unsold speaks to lost income, and selling that seat at any cost over the administration cost for a solitary traveler would have been an increasingly best situation.

The purpose of this project is to think about how aircraft ticket costs change after some time, remove the elements that impact these variances, and depict how they're associated (basically surmise the models that air bearers use to value their tickets). At that point, utilizing that data, construct a framework that can help buyers settle on buying choices by anticipating how air ticket costs will advance later on.

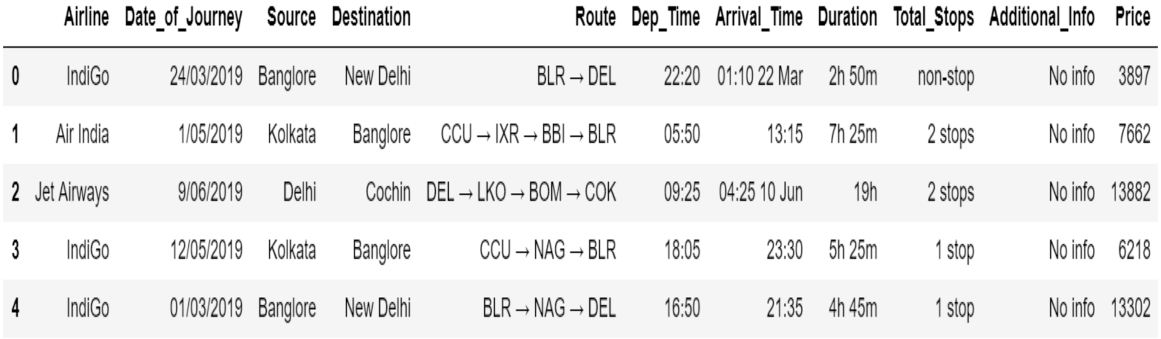
Related Work

The different research groups have concentrated on, for the most part, various arrangements of a different set of features and trained their models on various types of flights. A significant difference among these undertakings is the particular pattern they are attempting to foresee. Groves et al. proposed a model to foresee the anticipated least cost of all flights on a specific course. The model was moreover used to foresee cost with various objective properties, for example, the prediction from a particular flight, non-stop flight and so forth. Rama-Murthy built a model to anticipate the airfare cost with an explicit spotlight on how various elements impact the cost of aircraft tickets. Papadakis considered how costs of aircraft tickets change additional time by separating a few factors that conceivably influence the ticket value change and discovering their correlation.

Data Details

The collection of information is the most significant part of this project. There are different sources of information on various sites that are utilized to prepare the models. These websites give data about the numerous routes, times, carriers and fares. Different sources from APIs to buyer travel sites are accessible for information scraping.

There are sure features that would be extremely helpful in training a prediction algorithm, yet which the air-carriers, because of their competitive condition, don't discharge to the general population. This incorporates the real number of accessible seats on a flight, the dispersion of ticket buys over the lifetime of passage and fine-grained marketing projections. Likewise, the openly accessible admission information does exclude specific sorts of tickets, as a consolidator and corporate tickets, which are consulted in private with the aircraft. At long last, some ease aircraft don't distribute their rates on the booking frameworks, and are accordingly not filed by most online frameworks. Thus, for this project we have gathered dataset from Analytics India website which consist data of air carriers from India.



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

Fig 1. Training Data

Feature Selection

We have a dataset of flight details with eleven different features. The size of the training dataset is 10,683 records. At first we are going to train the models using this training dataset. After this, we are going to test the models which perform well with a test dataset that has 2671 records.

Features we have:

- **Airline**: The name of the airline. Depending on the company of the airline the fare prices changes.

- **Date of Journey**: Date of journey is essential feature to predict the flight ticket value as in peak period many people prefer to travel for example in Christmas vacation people travel to different part of the world

- **Source**: Location from where flight is departing is determining factor to predict fare

- **Destination**: The destination where the service ends

- **Route**: The route was taken by the flight to reach the destination.

- **Dep Time**: The time when the journey starts from the source.

- **Arrival Time**: Time of arrival at the destination.

- **Duration**: Total duration of the flight. This is one the important feature as based on time duration flight tickets changes. For example for smaller duration of flight tickets are cheaper and vice versa.

- **Total Stops**: Total stops between the source and destination. Lesser the stops higher the price.

- **Price**: The price of the ticket.

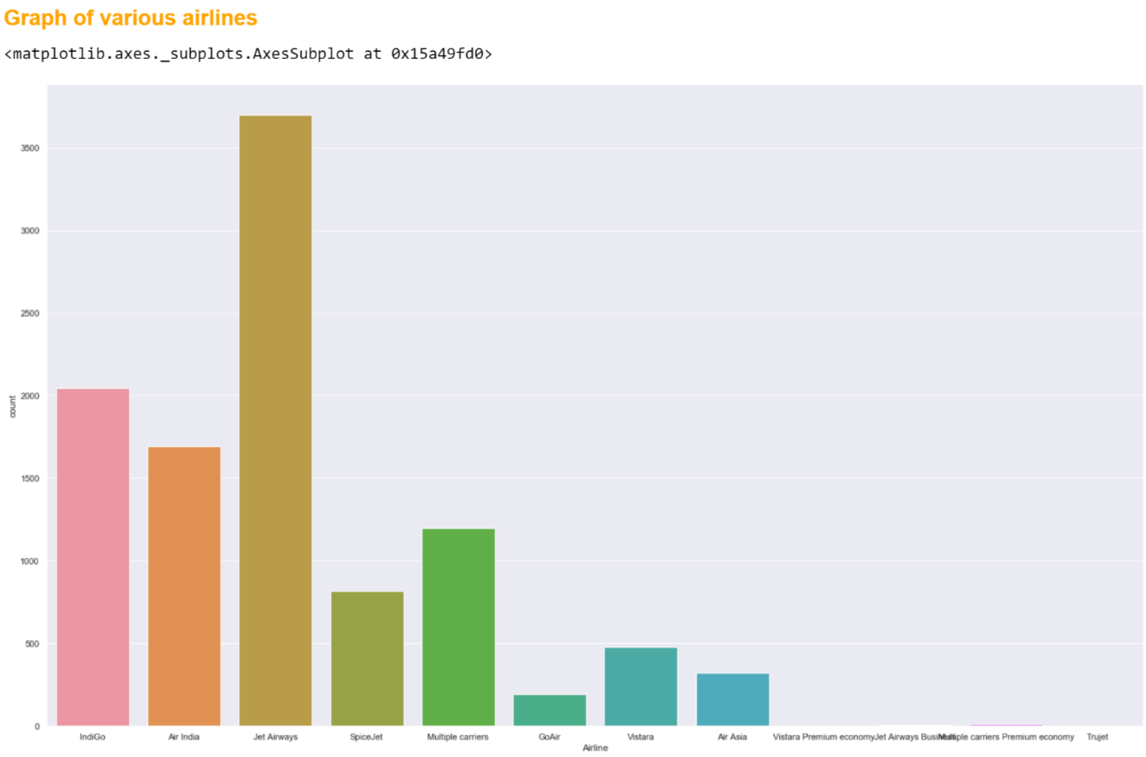
Milestone

1. Exploratory Data Analysis (EDA):

EDA is an approach for data analysis that employs a variety of techniques to maximize insight into a data set:

- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions
- Develop appropriate model
- Determine optimal factor settings

We observed the pattern of the data to perform Exploratory data analysis. It is important to analyse the data in order to find the outliers and anomalies in the data. While observing the data we looked at various features and its association with the other features.



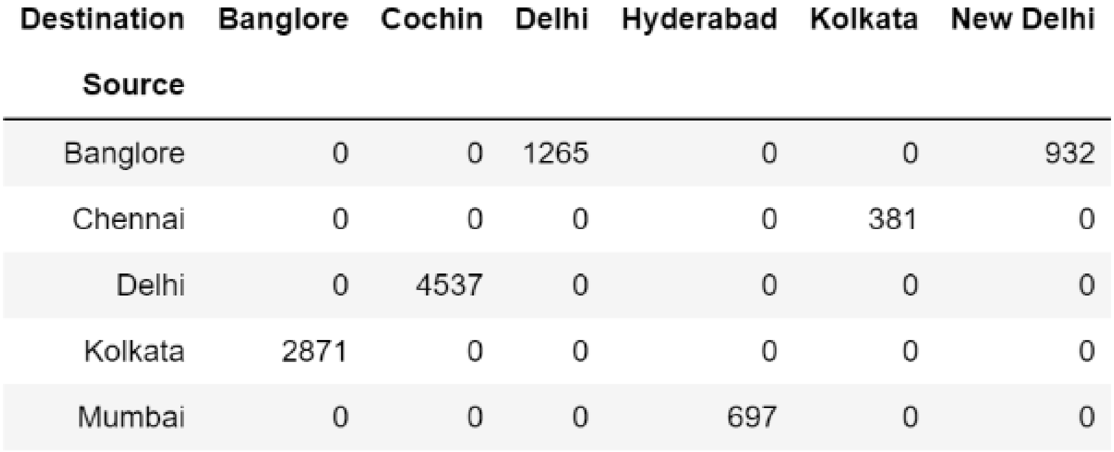
Graph of various airlines

```
<matplotlib.axes._subplots.AxesSubplot at 0x15a49fd0>
```

Fig 2. Number of Passengers vs Airline Company

Above figure gives us the insight of the number of passengers flying with particular airline company We can see that Jet airways has highest number of passengers travelling. IndiGo comes after that. On the other hand Vistara airlines has least number of passenge

Table 1 represents the number of passengers travelling from source city to destination city. From the table it is clear that the passengers are travelling from Delhi to Cochin more frequent comparative to other cities. On the other hand, Chennai to Kolkata has least number of frequent passengers.



Destination	Banglore	Cochin	Delhi	Hyderabad	Kolkata	New Delhi
Source						
Banglore	0	0	1265	0	0	932
Chennai	0	0	0	0	381	0
Delhi	0	4537	0	0	0	0
Kolkata	2871	0	0	0	0	0
Mumbai	0	0	0	697	0	0

Table 1. number of passengers from Source to destination

Following figure represents the number of passengers verses the number of total stops. People are majorly preferring the flights with less number of stops hence, we can see that one stop flights are being preferred more than two, three or four stop flights.

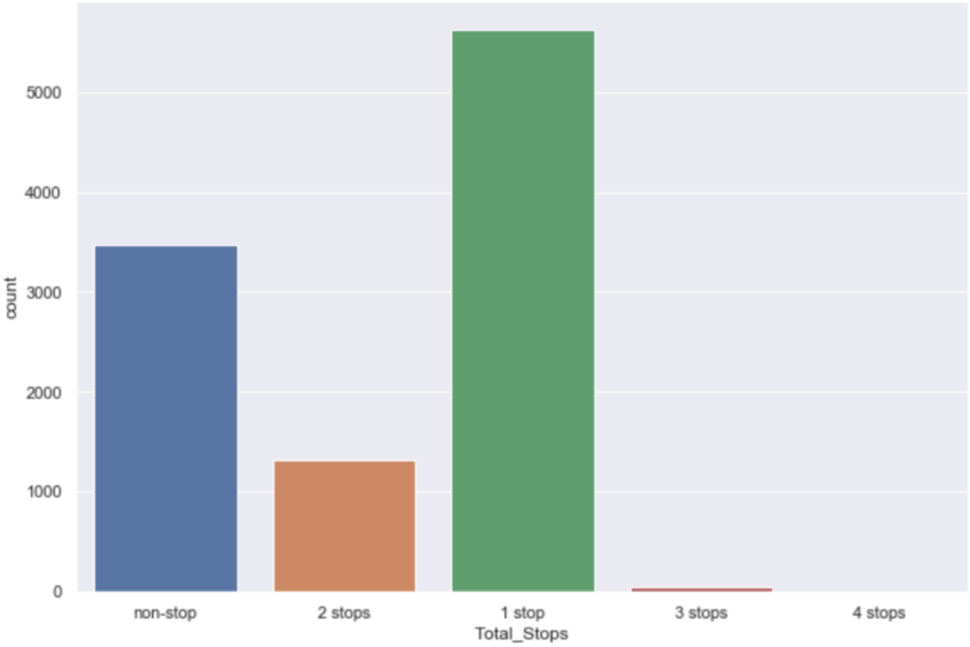


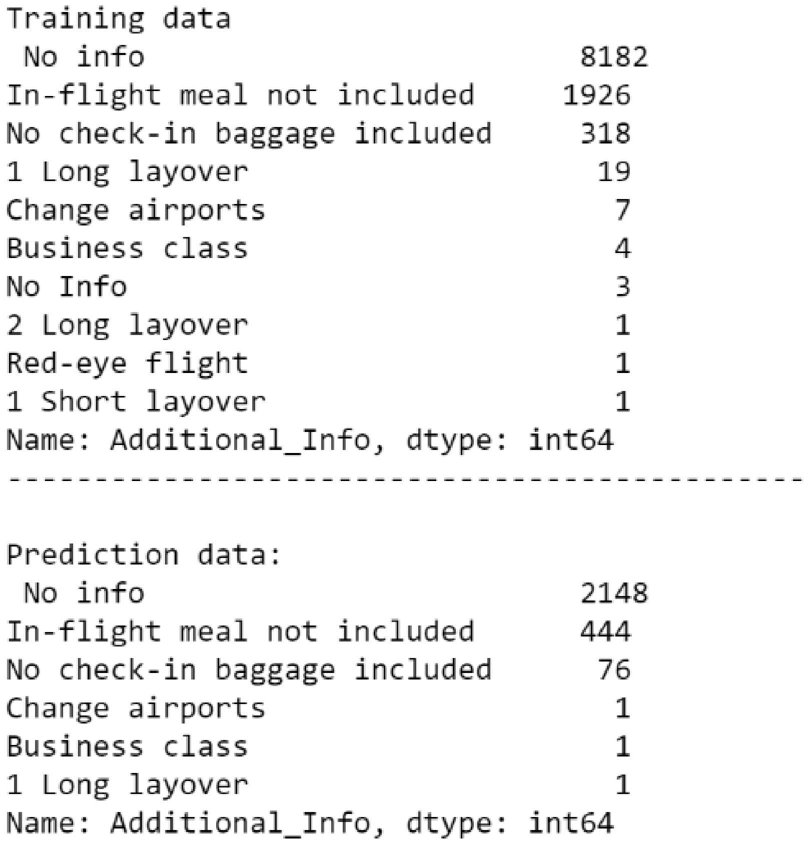
Fig 3. Number of stops vs number of flights

2. Transforming Data

All the collected data needed a lot of work so after the collection of data, it is needed to be clean and prepare according to the model requirements. All the unnecessary data is removed like duplicates and null values. In all machine learning models, this is the most important and time consuming step. Various statistical techniques and logic built in python libraries are used to clean and prepare the data. For example, Some entry in the dataset is a combination of both categorical and numerical. We have to perform a transformation to apply various models to this data.

We converted the format of the duration column given in the data. The format given was “hh:mm” which we converted into minutes to make the comparison between different flight durations easy.

Some of the entries in the training as well as Test data had null entries or had repeated entries. Following figure shows the entries of column “Additional information” from training data and test data. Here we can see that the “No info” and “No Info” is being counted two different values, even they just have the difference of capital and small character. Thus, we merged such anomalies.



```
Training data
 No info                         8182
In-flight meal not included     1926
No check-in baggage included     318
1 Long layover                    19
Change airports                    7
Business class                     4
No Info                            3
2 Long layover                     1
Red-eye flight                     1
1 Short layover                    1
Name: Additional_Info, dtype: int64
----------------------------------------------

Prediction data:
 No info                         2148
In-flight meal not included      444
No check-in baggage included      76
Change airports                    1
Business class                     1
1 Long layover                     1
Name: Additional_Info, dtype: int64
```

Fig 4. Entries of “additional information” column

In number of stops column we had inputs such as No stops, 1 stops, 2 stops, so for processing this data should be in numerical format hence, we converted that into 0, 1 , 2 respectively. Regarding the date of journey feature, we extracted the particular date, month and year separately for better processing.

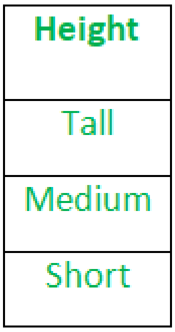
3. Categorial Data Encoding

Encoding data is also one of the important steps of data cleaning. Encoding is process of converting categorical data into numerical data so that it will be machine readable. Encoding also makes the processing easy. There are multiple types of encoding such as One hot encoding, Label encoding.

We have the data which is a combination of categorical and numerical data. So we divided the data into two sections based on type of the data. Then we converted the categorical data into numerical data using Label Encoder.

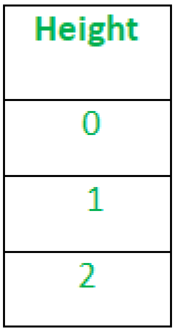
Label Encoder refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning. Following is the example of label encoding:

Suppose we have a column *Height* in some dataset.



Height
Tall
Medium
Short

After applying label encoding, the Height column is converted into:



Height
0
1
2

where 0 is the label for tall, 1 is the label for medium and 2 is label for short height.

4. Implementing Models

To develop the model for the flight price prediction, many machine learning algorithms are evaluated. All these models are implemented in the scikit learn. To evaluate the performance of these models, certain parameters are considered. Considering the observations we have seen above, We are going to choose the following models for our project:

- Linear Regression:

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the predictions of Y when plotted as a function of X form a straight line.

Linear regression finds the straight line, called the least squares regression line or LSRL, that best represents observations in data set. Suppose *Y* is a dependent variable, and *X* is an independent variable. The regression line is:

$$Y = B_0 + B_1X$$

Linear regression for our data gave the following matrix of accuracy and error

R Squared accuracy	0.5619
Mean Squared Error	0.117
Root Mean Squared Error	0.342

- Bayesian Regression Model:

Bayesian linear regression is an approach to linear regression in which the statistical analysis is undertaken within the context of Bayesian inference. When the regression model has errors that have a normal distribution, and if a particular form of prior distribution is assumed, explicit results are available for the posterior probability distributions of the model's parameters.

Advantage of Bayesian linear regression is that it allows a fairly natural mechanism to survive insufficient data, or poor distributed data. It allows us to put a prior on the coefficients and on the noise so that in the absence of data, the priors can take over.

Bayesian regression for our data gave the following matrix of accuracy and error

R Squared accuracy	0.561
Mean Squared Error	0.117
Root Mean Squared Error	0.342

- Random Forest Regression:

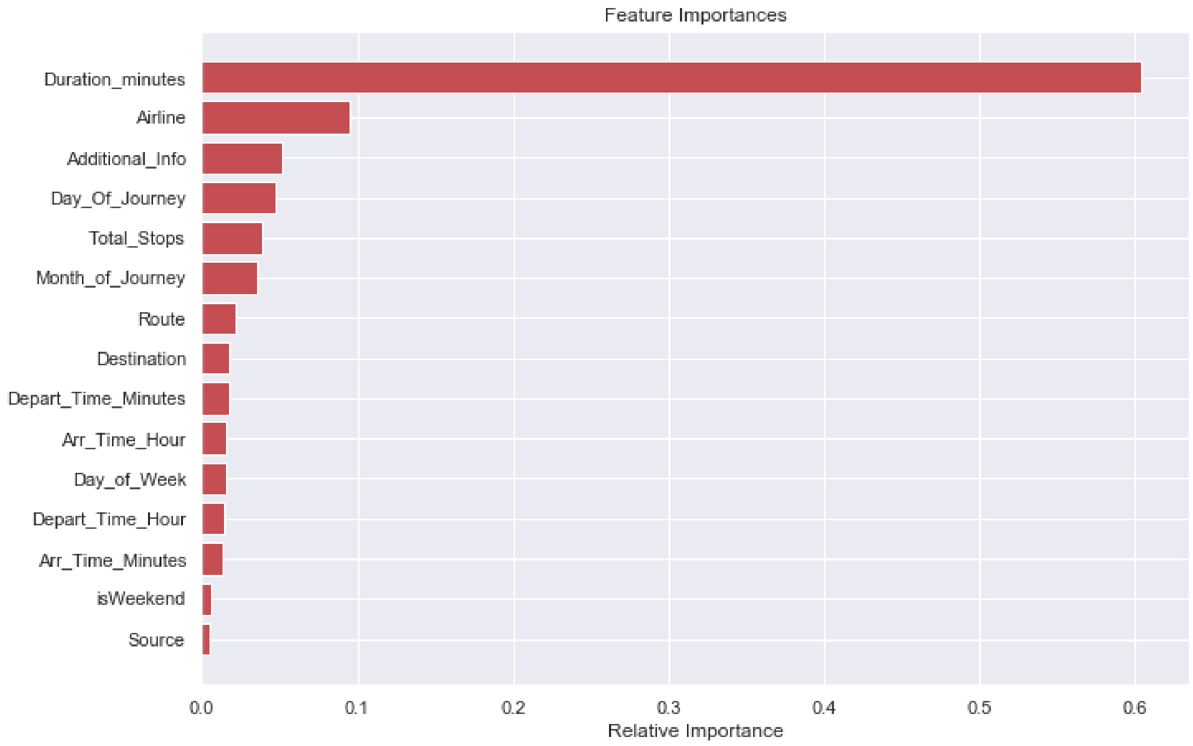
In random forests, trees are built from random samples of the training set. For each tree, the best split is chosen when splitting a node. This is summarized in the following pseudocode:

```
1 Randomly select k features from total m features, where k << m
2 Among the k features, calculate the node d using the best split point
3 Split the node into child nodes using the best split
4 Repeat steps 1-3 until l number of nodes has been reached
5 Repeat steps 1-4 n times to create n trees
```

Random Forest Regression for our data gave the following matrix of accuracy and error

R Squared accuracy	0.922
Mean Squared Error	0.020
Root Mean Squared Error	0.143

From Random forest regression we have got the important features. Feature importance is measured by observing how random re-shuffling (thus preserving the distribution of the variable) of each predictor influences model performance. Following figure gives us insights of the features importance. Duration of the flight is affecting the prediction in large amount than any other feature. Airline company is also important. On the other hand whether a day is weekday or weekend is least affecting the price.



- XGB Regression model:

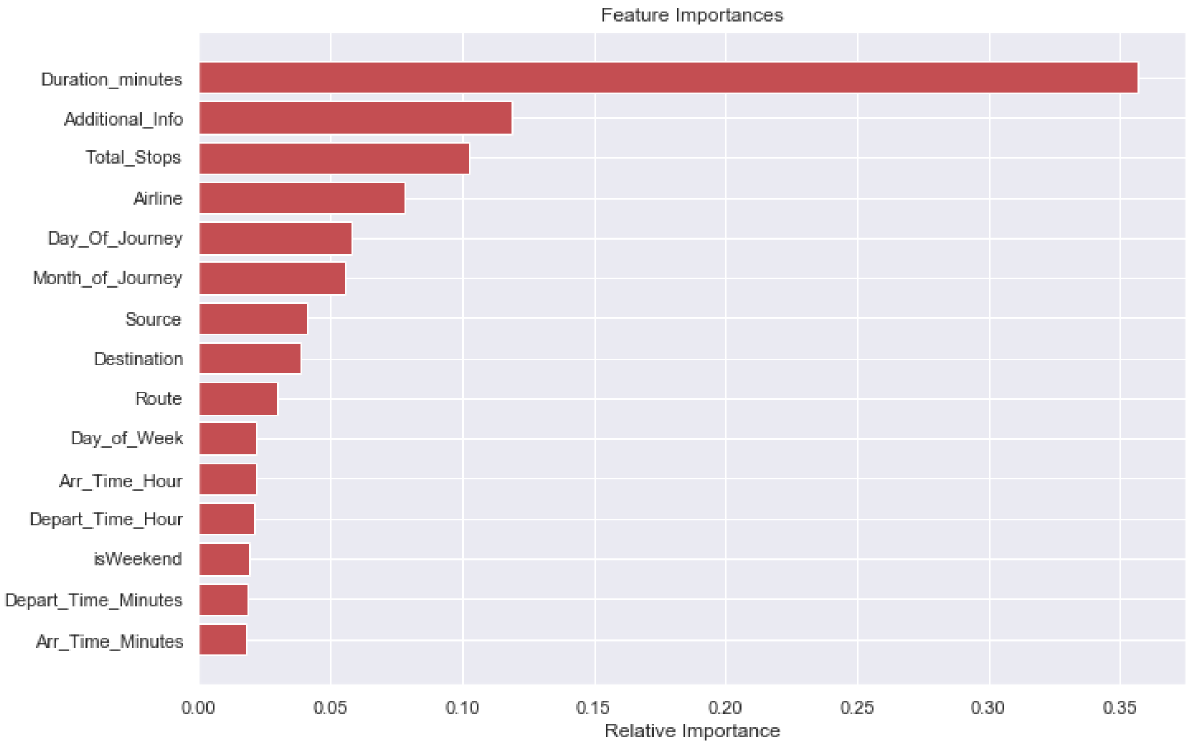
XGB regression is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. It is a perfect combination of software and hardware optimization techniques to yield superior results using less computing resources in the shortest amount of time.

XGB regression improves upon the base GBM framework through systems optimization and algorithmic enhancements. This model optimizes the systems using techniques such as Parallelization, Tree Pruning, Hardware Optimization.

XGB Regression for our data gave the following matrix of accuracy and error

R Squared accuracy	0.9367
Mean Squared Error	0.0166
Root Mean Squared Error	0.1290

Following figures give us the insights of the important features using XGB Regression model. Here we can see again duration is affecting the prediction which is followed by additional information.



- Gradient Boosting Regression Model:

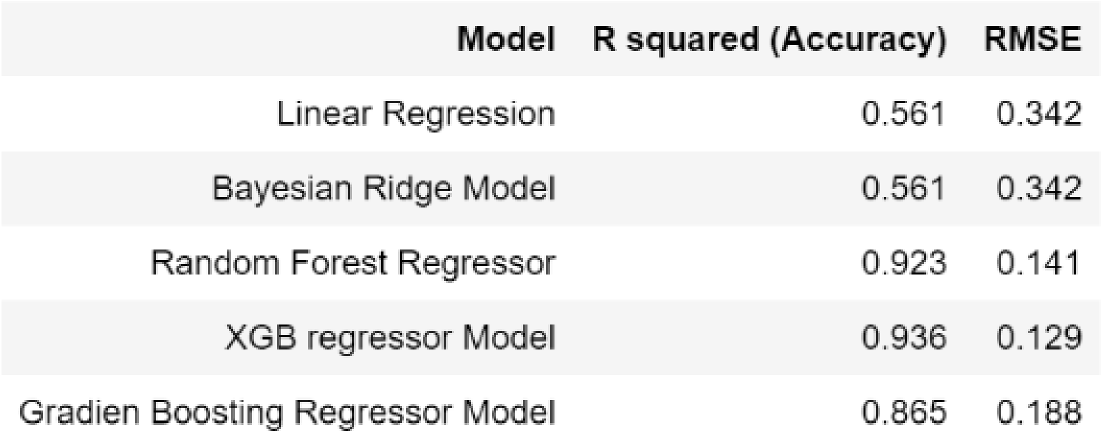
It is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Gradient Boosting Regression for our data gave the following matrix of accuracy and error

R Squared accuracy	0.865
Mean Squared Error	0.035
Root Mean Squared Error	0.188

5. Analysis using various_metrics

At last, we analyzed the outcomes of various models using multiple metrics. We used RMSE, R-squared, Mean Absolute error methods to analyse outcomes. The following table gives us the performance of various models we have implemented.



Model	R squared (Accuracy)	RMSE
Linear Regression	0.561	0.342
Bayesian Ridge Model	0.561	0.342
Random Forest Regressor	0.923	0.141
XGB regressor Model	0.936	0.129
Gradien Boosting Regressor Model	0.865	0.188

We can see that the accuracy of the XGB regressor is highest. Random Forest regression also does good job. Linear regression has lowest accuracy. Hence, we are going to choose two highest-performing models for the predicting the price of the flights. We are going to use the test data for this purpose on XGB regressor and Random Forest regression.

6. Applying the highest performing models on test data

We used two models XGB regressor and Random Forest regression on the test data. we were able to predict the prices of the flight tickets. We got 0.936 Accuracy score for XGB regressor and 0.92 R-squared accuracy for Random Forest regressor.

Conclusion

In this project, a machine learning framework was developed to predict the airfare prices. We utilized the flight dataset taken from AnalyticsIndia. Several features were extracted from the datasets and combined together with macroeconomic data, to implement in various machine learning models. With the help of the feature selection techniques, our proposed model is able to predict the airfare price with an adjusted R squared score of 0.936.