Implementation of Flight Fare Prediction System Using Machine Learning

ANURAG KUMAR JHA

DATE:08/11/2022

Abstract:

The Flight ticket prices increase or decrease every now and then depending on various factors like timing of the flights, destination, duration of flights. In the proposed system a predictive model will be created by applying machine learning algorithms to the collected historical data of flights. Optimal timing for airline ticket purchasing from the consumer's perspective is challenging principally because buyers have insufficient information for reasoning about future price movements. In this project we majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket. The project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like MumbaiDelhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum. The scope of the project can be extensively extended across the various routes to make significant savings on the purchase of flight prices across the Indian Domestic Airline market.

Keywords: Flight ticket, Optimal timing, historical data, competitive routes, Indian Domestic Airline market

PROBLEM STATEMENT:

Providing an approach for analysing the flight fare market for small and medium tourism companies. This ML model would help them know about future flight fares, increase/decrease in fare price over the future. This analysis would help small and medium tourism sector based companies maximise their profit as this ML model will help them predict best possible way for profit making by predicting flight fares between various cities across India and help them gather information accordingly.

Market/Customer Need Assessment:

This project aims to develop an application which will predict the flight prices for various flights using machine learning model. The user will get the predicted values and with its reference the user can decide to book their tickets accordingly. In the current day scenario flight companies try to manipulate the flight ticket prices to maximize their profits. There are many people who travel regularly through flights and so they have an idea about the best time to book cheap tickets. But there are also many people who are inexperienced in booking tickets and end up falling in discount traps made by the companies where actually they end up spending more than they should have. The proposed system can help save millions of rupees of customers by proving them the information to book tickets at the right time.

3.Target Specification and characterization:

Finding the best possible flight fare based on trends and parameters that have been used by the Machine Learning Algorithm .

Flight fare prediction will help both the customers as well as the small/medium tourism companies as affordable prices and good timings of flight fare booking will lead to more influx of customers who would want to book flights from companies booking affordable and timely flights by predicting situations ahead in the future.

Parameters used in the ML model are :

1) Size of Test Set: 10683 rows & 11 columns

2) Airline: The name of the airline.

3) Date of Journey: The date of the journey.

4) Source: The source from which the service begins.

5) Route: Route of the flight, start to end.

6) Destinations: The destination where the service ends.

7) Departure Time: The time when the journey starts from the source.

8) Arrival Time: Time of arrival at the destination.

9) Duration: Total duration of the flight.

10) Total Stops: Total stops between the source and destination.

11) Additional Info: Additional information about the flight

12) Price: The price of the ticket

4. External Search(information sources):

The dataset can be found on Kaggle . The dataset consists of airline name, Date of Journey, Source, Destination, Route , Departure time , Additional Info , Duration , Price and all.

The link for the data set: Flight Fare Prediction MH | Kaggle

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

        sns.set()
```

```
In [2]: train_data = pd.read_excel(r"Data_Train.xlsx")
```

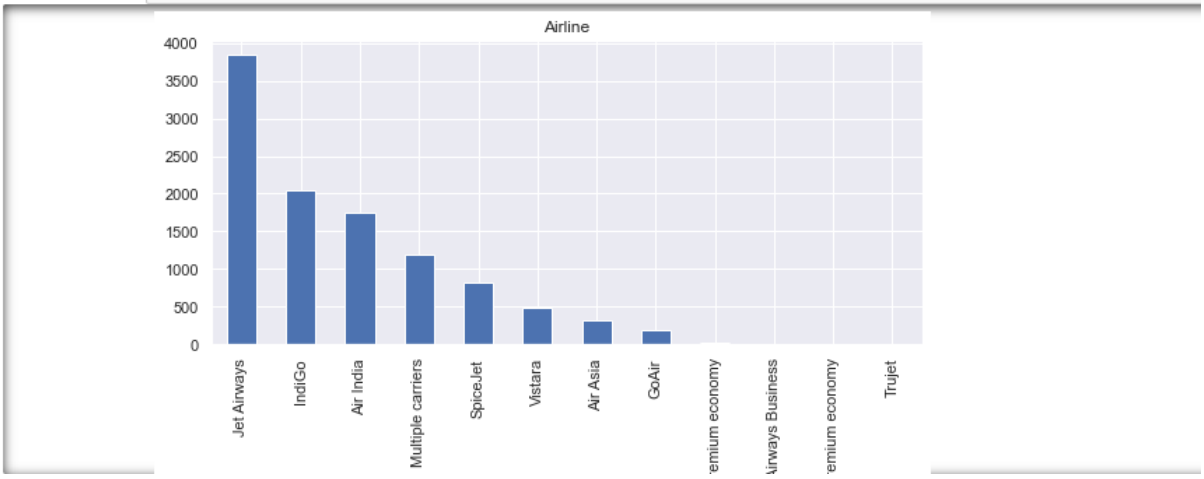
```
]: train_data.head()
```

]:

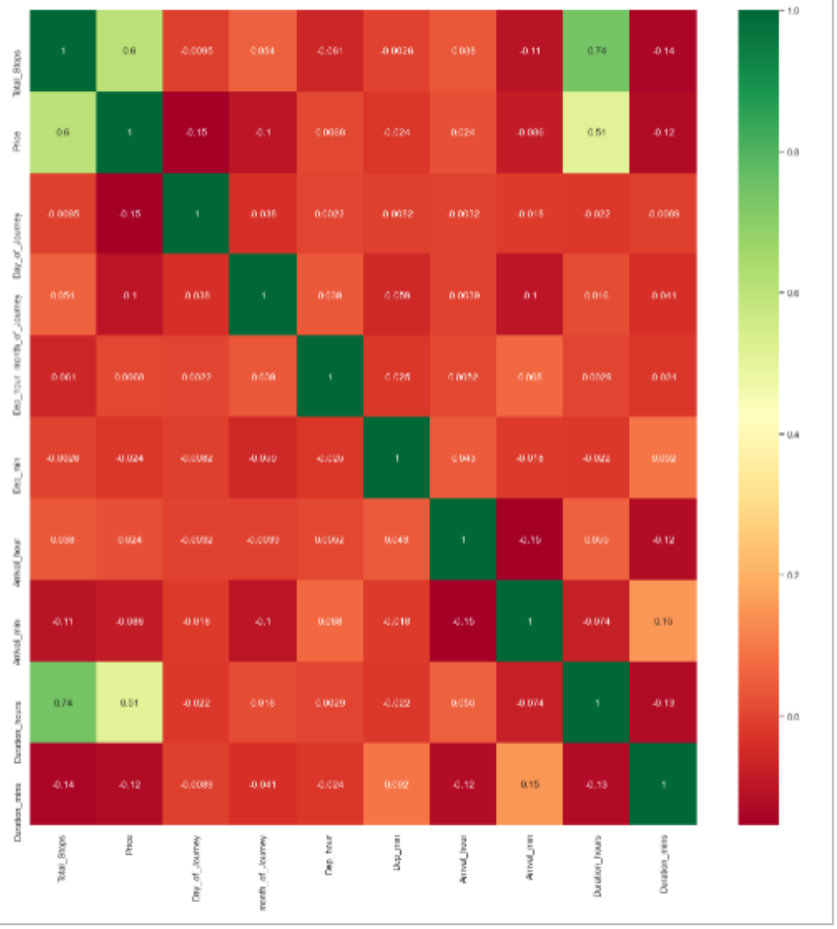
	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No 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

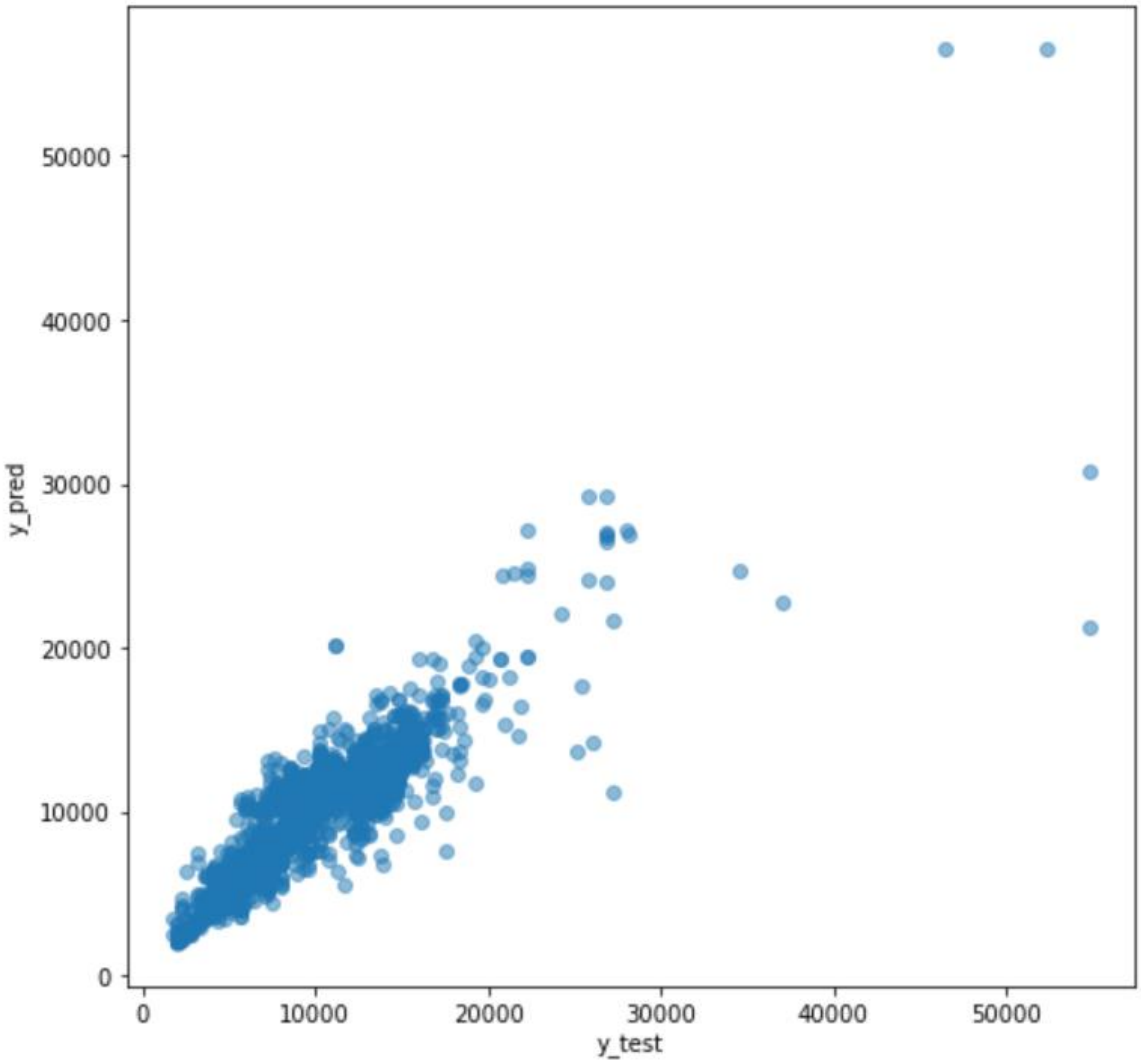
5.BENCHMARKING:

```
In [8]: #univariate analysis
        # Selecting the categorical columns
        categorical_col = train_data.select_dtypes(include=['object']).columns

        # Plotting a bar chart for each of the cateorical variable
        for column in categorical_col:
            plt.figure(figsize=(20,4))
            plt.subplot(121)
            train_data[column].value_counts().plot(kind='bar')
            plt.title(column)
```



```
plt.figure(figsize = (18,18))
sns.heatmap(train_data.corr(), annot = True, cmap = "RdYlGn")

plt.show()
```





The above data gives us a glance about the different types of flight companies along with their difference in numbers in the market prices and all along with the different source destination that the people chose in their journey as well as the weather, date and other factors that may contribute to increased prices in the use case.

6.Applicable Patents:

International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue V May 2022--The current patent may incorporate this patent for the inspiration of the methodology used as well as EDA analysis to some extent.

7. Applicable Regulations:

The patents mentioned above might claim the technology used if the algorithms are not developed and optimised individually and for our requirements. Using a pre-existing model is off the table if it incurs a patent claim.

1. Must provide access to the 3rd party websites to audit and monitor the authenticity and behaviour of the service.
2. Enabling open-source, academic and research community to audit the Algorithms and research on the efficacy of the product.
3. Laws controlling data collection: Some websites might have a policy against collecting customer data in form of reviews and ratings.
4. Must be responsible with the scraped data: It is quintessential to protect the privacy and intention with which the data was extracted.

8. Applicable Constraints:

1. The use of cloud platforms to store the data gathered over the net.
2. Using the spark service to clean and transform data .
3. For modelling using Timeseries and various models like Decision Tree , Linear Regression , Random Forest and all are tried and tested out for the best final model results .

9.Business Opportunity:

Currently, there are many fields where prediction-based services are used such as stock price predictor tools used by stock brokers and service like Zestimate which gives the estimated value of house prices. Therefore, there is requirement for service like this in the tourism and aviation industry which can help the customers in booking tickets while also leading in a significant boom in the tourism sector where both the customer and provider side will be able to make good profit . There are many research works that have been done on this using various techniques and more research is needed to improve the accuracy of the prediction by using different algorithms. More accurate data with better features can always be also be used to get more accurate results .

10.Concept Generation:

Motivation of coming up with the plan is to help people who tends to pay more for the flight fare ticket and for those who are naïve to this booking tickets process. This will also help us to get more exposure to the machine learning techniques that will help us to excel and improve in the existing skills.
The product/prediction system requires the tool of machine learning models to be written from scratch in order to suit our needs. . Tweaking these models for our use is less daunting than coding it up from scratch. A well trained model can either be repurposed or built. But building a model with the resources and data we have is dilatory but possible. The customer might want to spend the least amount of time giving input data. . This accuracy will take a little effort to nail, because it's imprudent to rely purely on Classic Machine Learning algorithm .

The Accuracy of the initial model is predicted by a code similar(PICTURE IS REPRESENTATIVE SIMILAR CODE IS USED FOR OUR DATASET)

```
model.fit(x_train, y_train,
          batch_size=batch_size,
          epochs=epochs,
          verbose=1,
          validation_data=(x_test, y_test))
score = model.evaluate(x_test, y_test, verbose=0)
print('Test loss:', score[0])
print('Test accuracy:', score[1])
```

```
x_train shape: (60000, 28, 28, 1)
60000 train samples
10000 test samples
Train on 60000 samples, validate on 10000 samples
Epoch 1/2
60000/60000 [==============================] - 6s 102us/sample - loss: 2.3050 - acc: 0.0848 - val_loss: 2.3045 -
val_acc: 0.0841
Epoch 2/2
60000/60000 [==============================] - 6s 103us/sample - loss: 2.3036 - acc: 0.0946 - val_loss: 2.3032 -
val_acc: 0.0925
Test loss: 2.3032085037231447
Test accuracy: 0.0925
```

11.Concept Development:

The concept can be developed by using the appropriate API (flask in this case) and using a GUI using Front End Framework. The cloud services has to be chosen accordingly to the need.

12.Final Project Prototype:

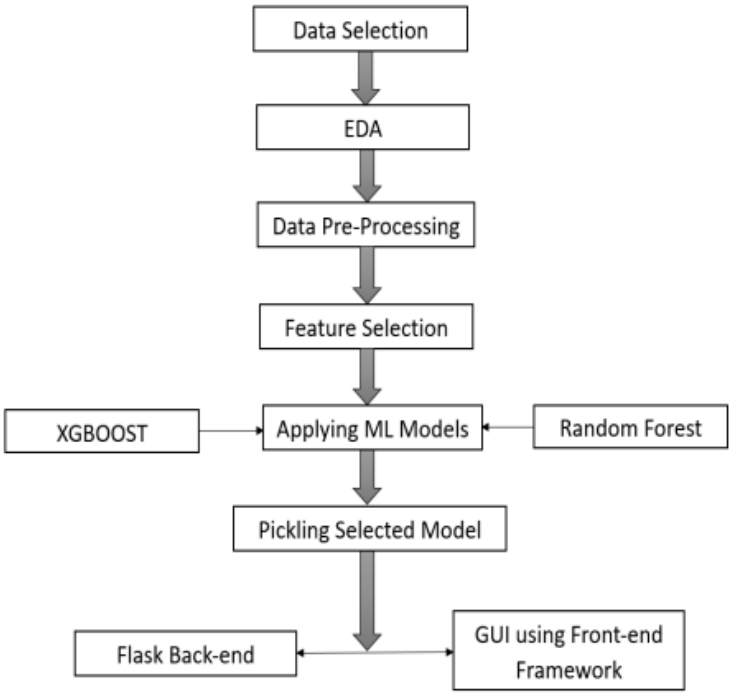


Fig. Proposed System Diagram

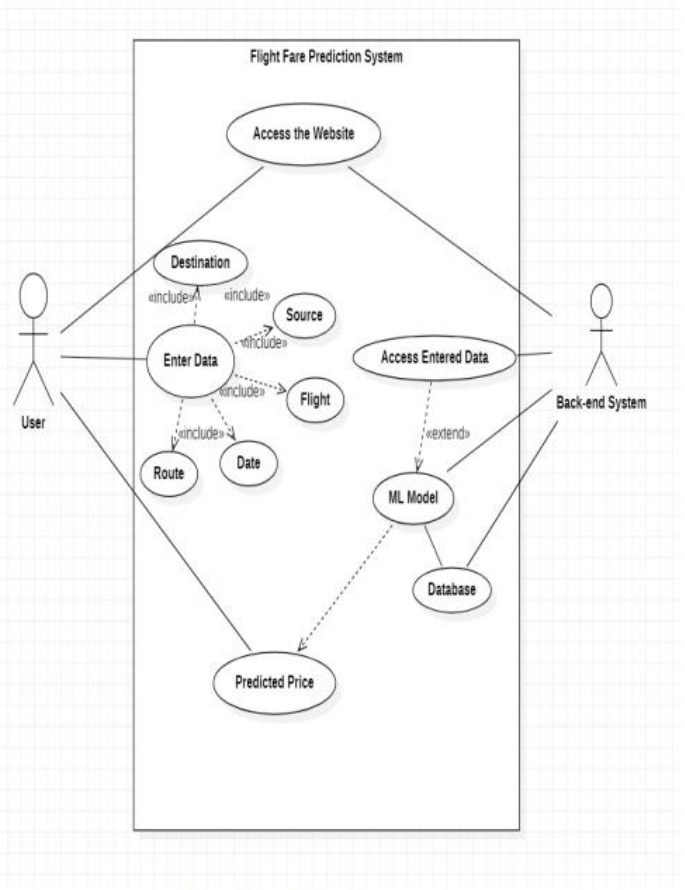


Fig. Use Case Diagram

The product takes the following functions to perfect and provide a good result.
Back-end:-
Model Development: This must be done before releasing the service. A lot of manual supervised machine learning must be performed to optimize the automated tasks.

0.Importing Necessary Libraries and reading the dataset using pandas .
1. Performing EDA to realize the dependent and independent features.
2.Handling Categorical Data and performing operation on test data EDA and Feature Engineering .
3. Algorithm training and optimization must be done to minimize overfitting of the model and hyperparameter tuning.
4.Feature Selection , Applying the most suitable ML algos and Pickling the File

Front End
1. Different user interface: The user must be given many options to choose form in terms of
parameters. This can only be optimized after a lot of testing and analysis all the edge
cases.
2. Interactive visualization the data extracted from the trained models will return raw and
inscrutable data. This must be present in an aesthetic and an “easy to read” style.
3. Feedback system: A valuable feedback system must be developed to understand the
customer’s needs that have not been met. This will help us train the models constantly.

13.Product Details:

An interactive user system will take inputs regarding the journey from the user and the user will get to know about the price for the journey that he wanted in real time considering the weather factor in mind regarding the same with the user interactive UI.
DIFFERENT ML ALGOS Tried Out:

A. Random Forest It is a supervised learning algorithm. The benefit of the random forest is, it very well may be utilized for both characterization and relapse issue which structure most of current machine learning framework. Random forest forms numerous decision trees, what’s more, adds them together to get an increasingly exact and stable expectation. Random Forest has nearly the equivalent parameters as a decision tree or a stowing classifier model. It is very simple to discover the significance of each element on the expectation when contrasted with others in this calculation.

The regular component in these techniques is, for the kth tree, a random vector theta k is produced, autonomous of the past random vector's theta 1, theta k-1 however with the equivalent distribution, while a tree is developed utilizing the preparation set and bringing about a classifier. x is an information vector. For a period, in stowing the random vector is created as the includes in N boxes where N is the number of models in the preparation set of information. In random split, choice includes various autonomous random whole numbers between 1 to K. The dimensionality and nature of theta rely upon its utilization in the development of a tree. After countless trees are created, they select the most famous class. These methodologies are called as random forests.

B. XGBoost XGboost is the implementation of gradient boosted decision tree. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all independent variables which are then fed into decision tree which predicts results. The weight of tree is predicted wrong by tree is increased then these variables are then fed to second decision tree. This individual classifiers/predictor then ensemble to give a strong and more precise model. It ca work on regression, classification, prediction, ranking, user-defined prediction problems.

C. Performance Metrics Performance metrics are statistical models which will be used to compare the accuracy of the machine learning models trained by different algorithms. The sklearn. metrics module will be used to implement the functions to measure the errors from each model using the regression metrics. Following metrics will be used to check the error measure of each model.

D. MAE (Mean Absolute Error) Mean Absolute Error is basically the sum of average of the absolute difference between the predicted and actual values. MAE = 1/n[∑(y-ý)] y = actual output values, ý = predicted output values n = Total number of data points Lesser the value of MAE the better the performance of your model.

E. MSE (Mean Square Error) Mean Square Error squares the difference of actual and predicted output values before summing them all instead of using the absolute value. MSE = 1/n[∑(y-ý)2] y=actual output values ý=predicted output values n = Total number of data points MSE punishes big errors as we are squaring the errors. Lower the value of MSE the better the performance of the model.

F. RMSE (Root Mean Square Error) RMSE is measured by taking the square root of the average of the squared difference between the prediction and the actual value. RMSE = √1/n[∑(y-ý)2] y=actual output values ý=predicted output values n = Total number of data points RMSE is greater than MAE and lesser the value of RMSE between different model the better the performance of that model.

G. R2 (Coefficient of Determination) It helps you to understand how well the independent variable adjusted with the variance in your model. R 2 = □ − ∑(ý-ȳ) 2 ÷ ∑(y-ȳ) 2 The value of R-square lies between 0 to 1. The closer its value to one, the better your model is when comparing with other model values.

14.CONCLUSION:

Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost. The values of R-squared obtained from the algorithm give the accuracy of the model. In the future, if more data could be accessed such as the current availability of seats, the predicted results will be more accurate. Finally, we conclude that this methodology is not preferred for performing this project. We can add more methods, more data for more accurate results