iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VI **Month of publication:** June 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.35053

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI Jun 2021- Available at www.ijraset.com*

# Prediction of Domestic Airline Tickets using Machine Learning

Pranita Rajure[1], Samiksha Bankar[2], Harsimran Bakshi[3], Bhakti Patil[4]
*[1, 2, 3]Student, [4]Faculty, Computer Department, AISSMSCOE*

***Abstract:*** ***Airlines usually keep their price strategies as commercial secrets and information is always asymmetric, it is difficult for ordinary customers to estimate future flight price changes. However, a reasonable prediction can help customers make decisions when to buy air tickets for a lower price. Flight price prediction can be regarded as a typical time series prediction problem. When you give customers a device that can help them save some money, they will pay you back with loyalty, which is priceless. Interesting fact: Fareboom users started spending twice as much time per session within a month of the release of an airfare price forecasting feature. Considering the features such as departure time, the number of days left for departure and time of the day it will give the best time to buy the ticket. Features are extracted from the collected data to apply Random Forest Machine Learning (ML) model. Then using this information, we are intended to build a system that can help buyers whether to buy a ticket or not. We have used Random Forest Algorithm which is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. With that said, random forests are a strong modelling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.***

***Keywords: Airline strategies, Airfare price prediction, Airfare changes, Random Forest algorithm, predictive model***

## I. INTRODUCTION

Since airlines usually keep their price strategies as commercial secrets and information is always asymmetric, it is difficult for ordinary customers to estimate future flight price changes [6]. However, a reasonable prediction can help customers make decisions when to buy air tickets for a lower price. Flight price prediction can be regarded as a typical time series prediction problem [6].

When you give customers a device that can help them save some money, they will pay you back with loyalty, which is priceless [11]. Interesting fact: **Fareboom** users started spending twice as much time per session within a month of the release of an airfare price forecasting feature [11].

We have used random forest algorithm which is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML [8]. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. With that said, random forests are a strong modelling technique and much more robust than a single decision tree. They aggregate many decision trees to limit over fitting as well as error due to bias and therefore yield useful results [9].

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [8]. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [8].

## II. LITERATURE SURVEY

This paper, gathered airfare data from a specific Greek airline corporation (Aegean Airlines) from the web and showed that it is feasible to predict prices for flights based on historical fare data and it uses Regression Tree Algorithm for prediction of airfare [2].

When you give customers advice that can help them save some money, they will pay you back with loyalty, which is priceless [11]. Interesting fact: **Fareboom** users started spending twice as much time per session within a month of the release of an airfare price forecasting feature. This tool continues to grow conversion for our partner [11].

In this paper, Random Forest Algorithm is used for prediction they combined two public datasets (DB1B and T-100) and predicts quarterly average airfare price with an adjusted R squared score of 0.869 and it uses Random Forest Algorithm for prediction [3].

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest [8]. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [8]. Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Y. Freund & R. Schapire, *Machine Learning*: *Proceedings of the Thirteenth International conference*, ***, 148–156), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression [8].

Since airlines usually keep their price strategies as commercial secrets and information is always asymmetric, it is difficult for ordinary customers to estimate future flight price changes [6]. However, a reasonable prediction can help customers make decisions when to buy air tickets for a lower price.

Flight price prediction can be regarded as a typical time series prediction problem [6].

This paper not only predicts the airfare but also identify which airfare's features that have the strongest impacts on the airfare changes [5]. It uses Random Forest and Multilayer Perceptron Algorithms [4].

The purpose of the paper is to study the factors which influence the fluctuations in the airfare prices and how they are related to the change in the prices [7]. Then using this information, build a system that can help buyers whether to buy a ticket or not [7].

In this article, we introduce a corresponding new command, rforest. We overview the random forest algorithm and illustrate its use with two examples: The first example is a classification problem that predicts whether a credit card holder will default on his or her debt [12]. The second example is a regression problem that predicts the log scaled number of shares of online news articles. We conclude with a discussion that summarizes key points demonstrated in the examples [12].

This paper proposes a novel application based on two public data sources in the domain of air transportation: the Airline Origin and Destination Survey (DB1B) and the Air Carrier Statistics database (T-100). The proposed framework combines the two databases, together with macroeconomic data, and uses machine learning algorithms to model the quarterly average ticket price based on different origin and destination pairs, as known as the market segment. The framework achieves a high prediction accuracy with 0.869 adjusted R squared score on the testing dataset [9].

In this tutorial, I will show you how to use Python to automatically surf a website like Expedia on an hourly basis looking for flights and sending you the best flight rate for a particular route you want every hour straight to your email [10].

## III. METHODOLOGY

### *A. Random Forest Algorithm*

Random Forest Algorithm *can* be used for both Classification and Regression problems in ML. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random-forest classifier:

*1)* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
*2)* The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm:

*a)* It takes less training time as compared to other algorithms.
*b)* It predicts output with high accuracy, even for the large dataset it runs efficiently.
*c)* It can also maintain accuracy when a large proportion of data is missing.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI Jun 2021- Available at www.ijraset.com*

The below diagram explains the working of the Random Forest algorithm:

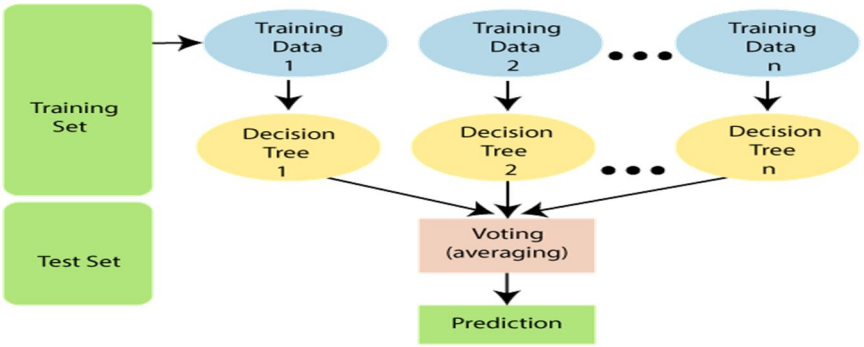



Fig 1 Random Forest Algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram shown above:

1) *Step-1:* Select random K data points from the training set.
2) *Step-2:* Build the decision trees associated with the selected data points (Subsets).
3) *Step-3:* Choose the number N for decision trees that you want to build.
4) *Step-4:* Repeat Step 1 & 2.
5) *Step-5:* For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

## IV. PROPOSED WORK

The aim of the implementation is to predict the flight ticket price using Random Forest (RF) ML Regression Algorithm. Earlier, many researchers have proposed a model for predicting the tickets based on various machine learning algorithms such as Linear Regression (LR), Support Vector Machines (SVMs), etc. Here we are using random forest algorithm because it gives a good accuracy as compared to other ML algorithms.

The major goals of our new system are as follows:

*A.* To design a system that can predict the flight fare so that a buyer can buy a ticket with the minimum fare
*B.* To provide lower price of tickets to buyers
*C.* To provide buyers best time to buy tickets.

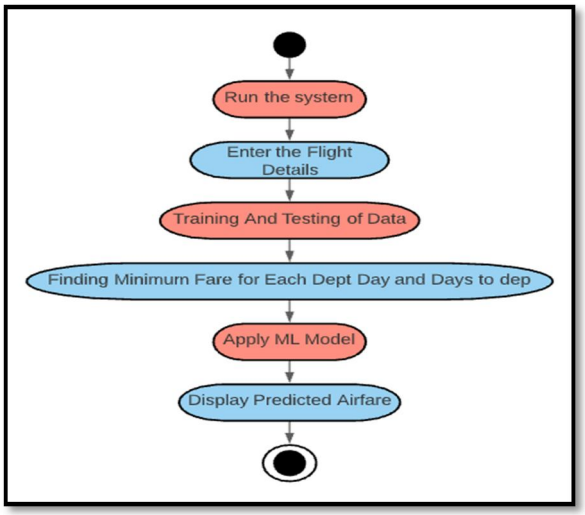



Fig 2 Proposed System Flow Diagram

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI Jun 2021- Available at www.ijraset.com*

The Architecture Diagram of our Model is given below. The Diagram shows sequence of processes that are used to build our model. There are five different steps to perform for model development. Start from collecting the required dataset for the system then, perform the remaining steps as shown in figure below.

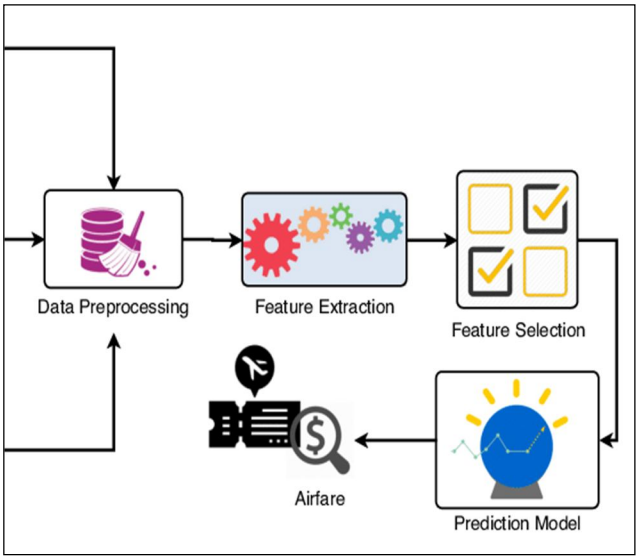



Fig 3 Architecture Diagram

## V. IMPLEMENTATION

### *A. Dataset*

We are using the dataset which consist of flight tickets for various airlines between the months of March and June of 2019 and between various cities. Size of training set: 10683 records and Size of test set: 2671 records.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Airline | e_of_Jour | Source | Destinatio | Route | Dep_Time | rrival_Tim | Duration | Total_Stop | ditional_In | Price |
| 2 | IndiGo | 24/03/201 | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 M | 2h 50m | non-stop | No info | 3897 |
| 3 | Air India | 1/05/2019 | Kolkata | Banglore | CCU → IXR | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 7662 |
| 4 | Jet Airway | 9/06/2019 | Delhi | Cochin | DEL → LKO | 09:25 | 04:25 10 J | 19h | 2 stops | No info | 13882 |
| 5 | IndiGo | 12/05/201 | Kolkata | Banglore | CCU → NA | 18:05 | 23:30 | 5h 25m | 1 stop | No info | 6218 |
| 6 | IndiGo | 01/03/201 | Banglore | New Delhi | BLR → NA | 16:50 | 21:35 | 4h 45m | 1 stop | No info | 13302 |
| 7 | SpiceJet | 24/06/201 | Kolkata | Banglore | CCU → BL | 09:00 | 11:25 | 2h 25m | non-stop | No info | 3873 |
| 8 | Jet Airway | 12/03/201 | Banglore | New Delhi | BLR → BO | 18:55 | 10:25 13 M | 15h 30m | 1 stop | In-flight m | 11087 |
| 9 | Jet Airway | 01/03/201 | Banglore | New Delhi | BLR → BO | 08:00 | 05:05 02 M | 21h 5m | 1 stop | No info | 22270 |
| 10 | Jet Airway | 12/03/201 | Banglore | New Delhi | BLR → BO | 08:55 | 10:25 13 M | 25h 30m | 1 stop | In-flight m | 11087 |
| 11 | Multiple ca | 27/05/201 | Delhi | Cochin | DEL → BO | 11:25 | 19:15 | 7h 50m | 1 stop | No info | 8625 |
| 12 | Air India | 1/06/2019 | Delhi | Cochin | DEL → BL | 09:45 | 23:00 | 13h 15m | 1 stop | No info | 8907 |
| 13 | IndiGo | 18/04/201 | Kolkata | Banglore | CCU → BL | 20:20 | 22:55 | 2h 35m | non-stop | No info | 4174 |
| 14 | Air India | 24/06/201 | Chennai | Kolkata | MAA → CC | 11:40 | 13:55 | 2h 15m | non-stop | No info | 4667 |
| 15 | Jet Airway | 9/05/2019 | Kolkata | Banglore | CCU → BO | 21:10 | 09:20 10 M | 12h 10m | 1 stop | In-flight m | 9663 |
| 16 | IndiGo | 24/04/201 | Kolkata | Banglore | CCU → BL | 17:15 | 19:50 | 2h 35m | non-stop | No info | 4804 |
| 17 | Air India | 3/03/2019 | Delhi | Cochin | DEL → AM | 16:40 | 19:15 04 M | 26h 35m | 2 stops | No info | 14011 |
| 18 | SpiceJet | 15/04/201 | Delhi | Cochin | DEL → PN | 08:45 | 13:15 | 4h 30m | 1 stop | No info | 5830 |
| 19 | Jet Airway | 12/06/201 | Delhi | Cochin | DEL → BO | 14:00 | 12:35 13 J | 22h 35m | 1 stop | In-flight m | 10262 |
| 20 | Air India | 12/06/201 | Delhi | Cochin | DEL → CC | 20:15 | 19:15 13 J | 23h | 2 stops | No info | 13381 |
| 21 | Jet Airway | 27/05/201 | Delhi | Cochin | DEL → BO | 16:00 | 12:35 28 M | 20h 35m | 1 stop | In-flight m | 12898 |

Sheet1

Fig 4 Dataset

*B. Data Pre-Processing*

Before building model, the data should be properly preprocessed and converted to quality, clean data even the resulting machine learning model will be of great quality.The data pre-processing includes three main parts that is data integration, data cleaning, data transformation. In data integration the data collected from various sources are integrated. In data cleaning process the data containing the null values, unnecessary rows with null values are being cleared. The data transformation includes the feature scaling, categorial data,etc to set the certain range of data.

A good data preprocessing in machine learning is the most important factor that can make a difference between a good model and a poor machine learning model. So, we need to do pre-process the data to get the perfect accuracy.

```
23 # As Airline is Nominal Categorical data we will perform OneHotEncoding

   Airline = train_data[["Airline"]]

   Airline = pd.get_dummies(Airline, drop_first= True)

   Airline.head()
```

23

| | Airline_Air India | Airline_GoAir | Airline_IndiGo | Airline_Jet Airways | Airline_Jet Airways Business | Airline_Multiple carriers | Airline_Multiple carriers Premium economy | Airline_SpiceJet | Airline_Trujet | Air |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Fig 5 Pre-processing

*C. Feature Extraction*

In this the features which are provided in data set are converted into the its easier form that is it means transforming raw data into a feature vector which helps in building the model. Making the data or the feature in its easier format would help to train the our model easily also it provides the a good accuracy rate.

```
9  train_data["Journey_day"] = pd.to_datetime(train_data.Date_of_Journey, format="%d/%m/%Y").dt.day

10 train_data["Journey_month"] = pd.to_datetime(train_data["Date_of_Journey"], format = "%d/%m/%Y").dt.month

11 train_data.head()
```

11

| y | Source | Destination | Route | Dep_Time | Arrival_Time | Duration | Total_Stops | Additional_Info | Price | Journey_day | Journey_month |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Banglore | New Delhi | BLR → DEL | 22:20 | 01:10 22 Mar | 2h 50m | non-stop | No info | 3897 | 24 | 3 |
| | Kolkata | Banglore | CCU → IXR → BBI → BLR | 05:50 | 13:15 | 7h 25m | 2 stops | No info | 7662 | 1 | 5 |
| | | | DEL → LKO | | | | | | | | |

Fig 6 Feature Extraction

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VI Jun 2021- Available at www.ijraset.com*

*D. Feature Selection*

In the feature selection, we can find out the best feature which will contribute and have good relation with target variable. Some of the feature selection methods are Heatmap, FeatureImportances, selectKbest. In this we can find out which feature of data is independent and dependent, also to find out which feature cointains the most importance we can do this by using the Extra tree Regressor.

```
48 # Important feature using ExtraTreesRegressor

   from sklearn.ensemble import ExtraTreesRegressor
   selection = ExtraTreesRegressor()
   selection.fit(X, y)

48 ExtraTreesRegressor(bootstrap=False, ccp_alpha=0.0, criterion='mse',
                       max_depth=None, max_features='auto', max_leaf_nodes=None,
                       max_samples=None, min_impurity_decrease=0.0,
                       min_impurity_split=None, min_samples_leaf=1,
                       min_samples_split=2, min_weight_fraction_leaf=0.0,
                       n_estimators=100, n_jobs=None, oob_score=False,
                       random_state=None, verbose=0, warm_start=False)

49 print(selection.feature_importances_)

   [1.95434120e-01 1.44021586e-01 5.38555703e-02 2.40539410e-02
    2.17430940e-02 2.73470116e-02 1.96095029e-02 1.27500039e-01
    1.74337150e-02 1.06741541e-02 1.87128973e-03 1.65718507e-02
    1.51364596e-01 6.81637760e-02 1.89902192e-02 8.47326514e-04
    3.10615675e-03 1.08644888e-04 5.29446496e-03 9.35369428e-05
    6.27777871e-04 1.23137157e-02 3.24867396e-03 8.91085147e-03
    1.41174133e-02 1.95895945e-02 7.68279226e-03 4.01014841e-04
    2.50235698e-02]
```

Fig 7 Feature selection

Following shows the visual representation of feature's importance

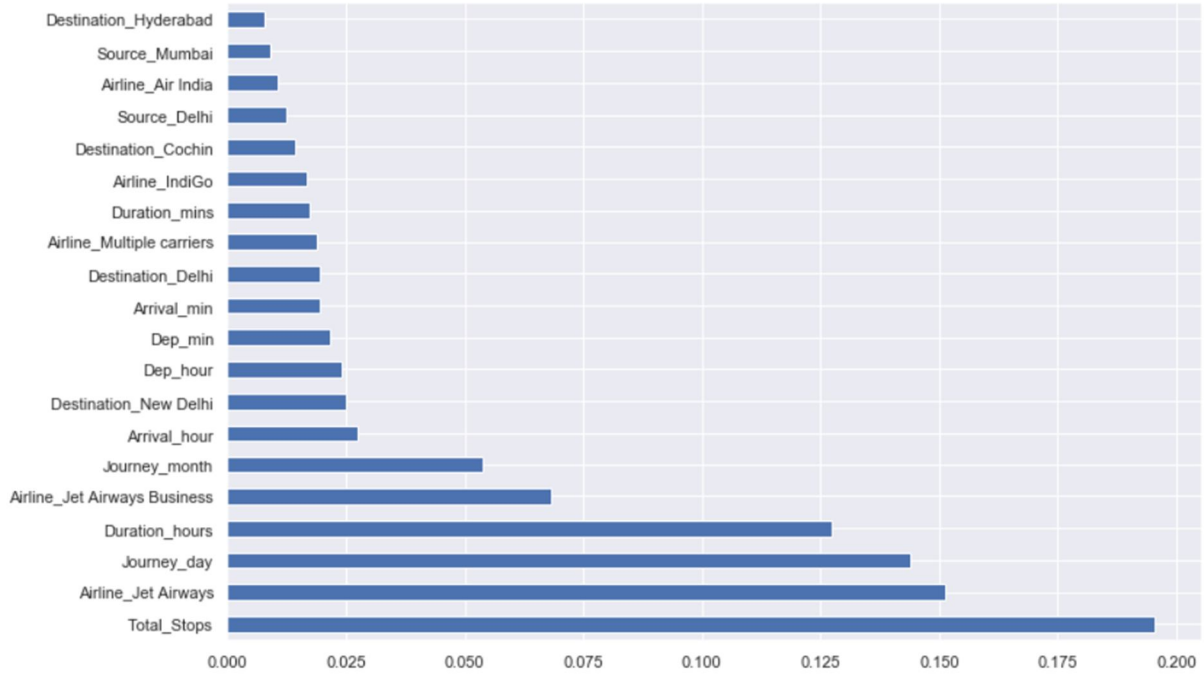



Fig 8 Feature's importance

*E. Fitting the Model*

For fitting of the model we split dataset into train and test set in order to prediction. After splitting, you will train the model on the training set and perform the predictions on the test set. After training, check the accuracy using actual and predicted values. Once the training of the model is done, we can store that model using pickle file so that we can reuse it again. Thus, you can predict the price.

```
51 from sklearn.model_selection import train_test_split
   X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)

52 from sklearn.ensemble import RandomForestRegressor
   reg_rf = RandomForestRegressor()
   reg_rf.fit(X_train, y_train)

52 RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                         max_depth=None, max_features='auto', max_leaf_nodes=None,
                         max_samples=None, min_impurity_decrease=0.0,
                         min_impurity_split=None, min_samples_leaf=1,
```

Fig 9 Training of model

```
62 import pickle
   # open a file, where you ant to store the data
   file = open('flight_rf.pkl', 'wb')

   # dump information to that file
   pickle.dump(reg_rf, file)

63 model = open('flight_price_rf.pkl','rb')
   forest = pickle.load(model)

64 y_prediction = forest.predict(X_test)

65 metrics.r2_score(y_test, y_prediction)
65 0.8117443234361064
```

Fig 10 Fitting the Model into Pickle File

## VI. RESULTS

Following shows the output of our system:

*A. This is the UI of our Project*


Fig 11 UI for system

*B. In this user have to enter valid details like destination, source, etc to get the flight ticket price.*

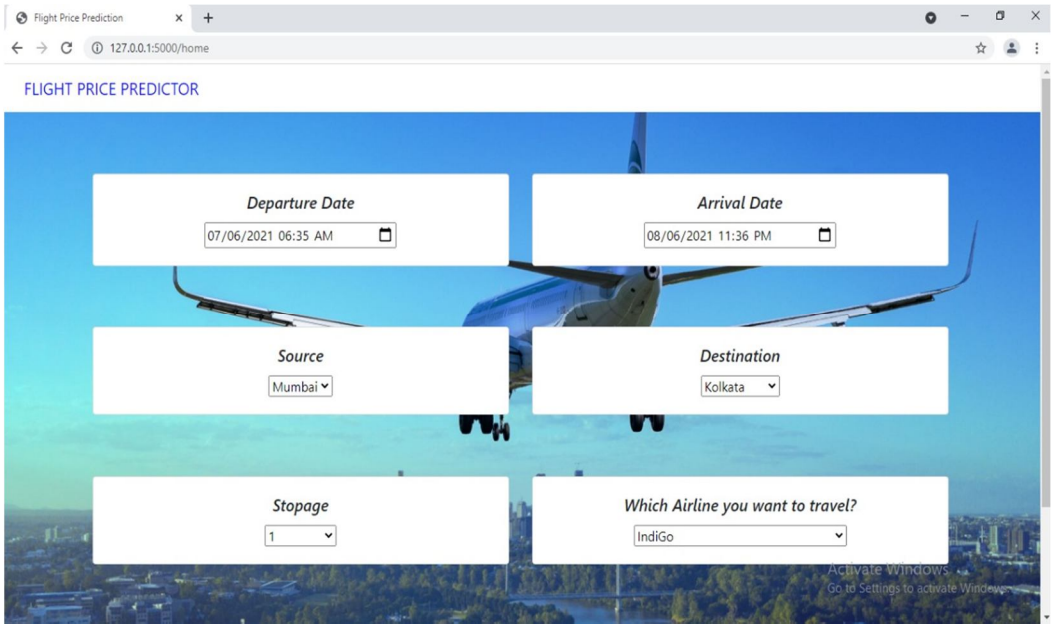



Fig 12 UI for system

*C. After entering the details, the user will get to know the predicted flight ticket price for a particular destination and source for a particular time or date according to the user's convienence.*

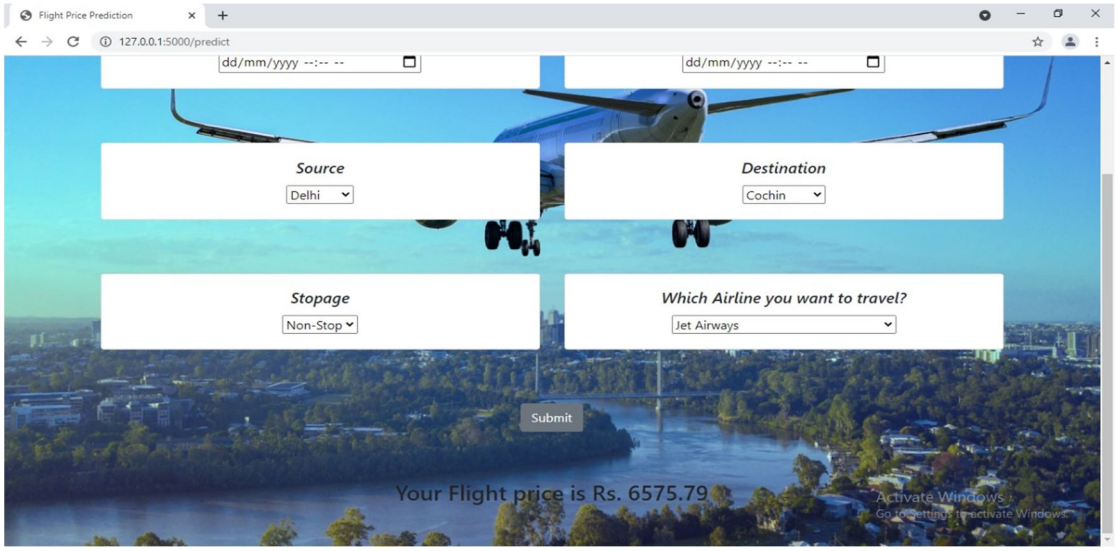



Fig 13. Final output

## VII. CONCLUSIONS

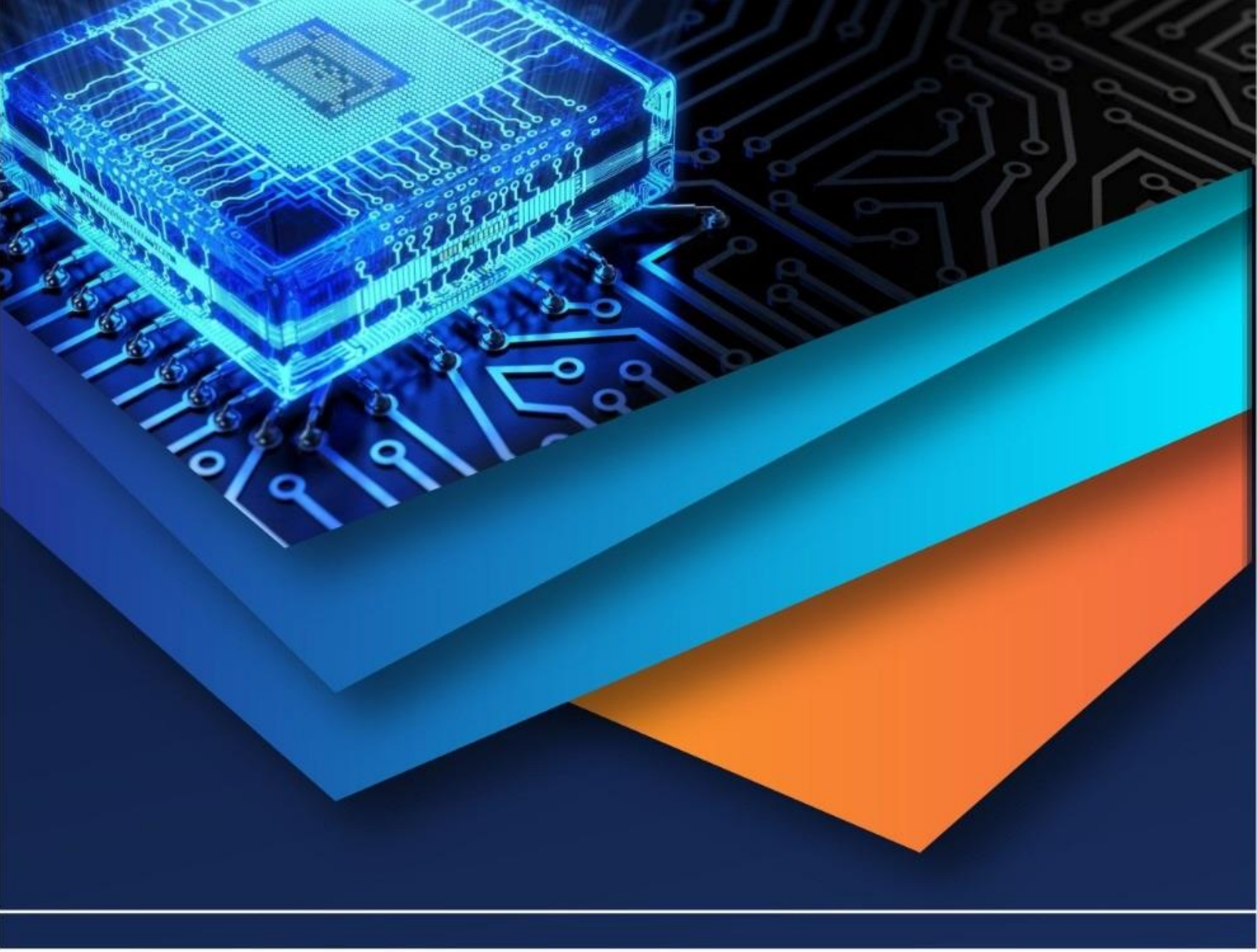
The various machine learning algorithms such as Linear Regression (LR), Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbors, etc are used in the previous systems for predicting the flight ticket price. In our ML based system, we used the Random Forest Algorithm which gives more accuracy in predicting the airfare. Features such as departure time, the number of days left for departure and time of the day, etc will be used for predicting the flight ticket price.

This system also helps the customers to buy the flight ticket at lower price. The system is easy to use and it gives more accuracy in prediction. Also, the time required for prediction is less and which helps the customers to get price quickly. This saves the time of the customers. Also, getting the flight ticket price in advance will help the customer in decision making whether to buy the ticket or not according to their convenience.

Hence our model gives the accuracy of 81% for prediction. We implemented Hyperparameter Tuning to increase the accuracy of our model. We have created a web application for predicting the airfare. In future we will implement mobile application for the same and we will add other solutions such as Hotels, Airport conditions, etc… In future, we will collect more data to increase the accuracy of our model.

## REFERENCES

[1] Pranita Rajure, Samiksha Bankar, Harsimran Bakshi and Bhakti Patil, "A Machine Learning Approach to Predict Price of Airlines Tickets", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 9, Issue II (Feb 2021)
[2] K. Tziridis, Th. Kalampokas, G.A. Papakostas, "Airfare prices prediction using machine learning techniques" IEEE,2017.
[3] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Yudong Tao, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen, "A Framework for Airfare Price Prediction: A Machine Learning Approach." IEEE(IRI),2019.
[4] Abhilash1, Ranjana Y2, Shilpa S3, Zubeda A Khan, "Survey on Air Price Prediction using Machine Learning Algorithms", IJIREEICE, 2019.
[5] Viet Hoang Vu; Quang Tran Minh; Phu H. Phung, "An airfare prediction model for developing markets",IEEE,2018
[6] T. Liu, J. Cao, Y. Tan, and Q. Xiao, "ACER: An adaptive context-aware ensemble regression model for airfare price prediction," in the international conference on progress in informatics and computing, 2017, pp. 312–317.
[7] Supriya Rajankar; Neha Sakharkar, "A Survey on Flight Pricing Prediction using Machine Learning", IJERT,2019
[8] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, October 2001.
[9] Predicting_Flight_Prices_in_India, journel, researchgate publication, 33782141
[10] https://dzone.com/articles/make-python-surf-the-web-for-you-and-send-best-fli
[11] https://www.altexsoft.com/blog/business/price-forecasting-machine-learning-based-approaches-applied-to-electricity-flights-hotels-real-estate-and-stock-pricing/
[12] https://journals.sagepub.com/doi/10.1177/1536867X20909688



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY