# Flight Price Prediction Using Machine Learning

[1]Victor Sarmacharjee, [2]Himangshu Nath, [3]Ashim Buragohain, [4]Rajesh Kumar Gouda, [5]Dibya Jyoti Bora
[1]Department of Information Technology, SCS, The Assam Kaziranga University, India
[2]Department of Information Technology, SCS, The Assam Kaziranga University, India
[3]Department of Information Technology, SCS, The Assam Kaziranga University, India
[4]Department of Information Technology, SCS, The Assam Kaziranga University, India
[5]Department of Information Technology, SCS, The Assam Kaziranga University, India
[1] cs18msiit042@kazirangauniversity.in, [2] cs18msiit007@kazirangauniversity.in,
[3] cs18msiit041@kazirangauniversity.in, [4] cs18msiit012@kazirangauniversity.in,
[5] dibyajyotibora@kzu.ac.in

## ABSTRACT

Those who frequently travel will be better educated about the best offers and the best times to buy tickets. For financial reasons, a lot of airlines change their prices according to the seasons or the time of year. The price will increase as more people travel. The real idea behind our travel prediction system is to forecast flight expenses by comparing today's pricing to yesterday's. Using various machine learning techniques on a sizable dataset, we will build a model to forecast flight prices, and the effectiveness of the models will be compared.

Keywords: Indian Airlines, Machine Learning, Exploratory Data Analysis, Prediction Model, Pricing Models, Model Training & Testing, Model Evaluation.

## INTRODUCTION

The key components of any contemporary transportation system are passenger airplanes, cargo airlines, and air traffic control systems. Countries from all around the world have tried to create various techniques over time to improve the aviation system. The way airlines function as a result has changed dramatically. Flight delays can be inconvenient for modern travelers. Around $45 billion in lost time and money is incurred by passengers each year as a result of almost 70% of airline flights being cancelled or delayed. Machine learning is the idea of learning through pattern recognition and experience accumulation without explicit programming. Training an ML model involves providing training data for the algorithm to use as a foundation for learning. Both organized and unstructured data can be categorized. The purpose of this study is to develop and improve models that can predict airfare far in advance and to gain a better understanding of the variables that affect airfare. While the client seeks the lowest price, the airlines' main objective is to turn a profit. The aim of this research is to create and enhance models that can forecast airfare months in advance and to better comprehend the factors that influence airfare. Although the customer is looking for the best deal, the airlines' top priority is to make money.

## LITERATURE REVIEW

The authors of the proposed study used a dataset of 1775 data flights from Alliance Airlines for their research, and they used this data to develop a machine learning model in order to predict the cost of airline tickets. Several quantities of features were used to train the model to show how the features chosen could influence a model's accuracy. many strategies, such as the Generalized Regression. Random Forest Regressor and neural networks have been used. For each machine learning algorithm, Linear Regression (LR), Regression Tree, and SVM, different results were achieved. They have experimented with and trained a variety of models while removing and adding various features from the dataset. [1] For the purpose of calculating the price of aero plane tickets, several machine learning techniques have been developed. The algorithms are: Naive Bayes Regressor, Linear Regression, and X Boost Regressor. The X boost regressor method has the highest accuracy of all of these. [2] Finding a machine learning model that could more precisely forecast the price of flights to India was the goal of this project. After putting various models to the test, it was discovered that the Random Forest algorithm produced an accurate prediction of the outcome. In terms of results, the essay outperforms models, and it aspires to do even better in the future. [3] In their work on the Flight Fare Prediction System, researchers applied various machine learning algorithm techniques, including Random Forest, Decision tree, and Linear regression, to a dataset to determine the ideal time to book a ticket. The goal of this project is to create a machine learning-based programmed that estimates flight costs for various flights. They list the approaches they used as being Linear Regression, Decision Trees, and Random Forest. MAE, MSE, and RSME are the performance measurement techniques. The results of their experiment weren't fully accurate, but they will be more accurate if more real-time data sets are used. [4] Enhancing the ML model to predict the typical airline ticket price for business reasons. For predicting the mean plane price in our model while adjusting the R squared score, feature selection methods were recommended. comparing the results of various machine learning classifiers exposes the scope of the plane price prediction problem. [5] Authors have reported on the task of collecting plane data from a flying operator via their website and shown that it is possible to predict plane fares based on previously acquired data. The results demonstrate that ML models can accurately forecast airline costs, and a number of additional aspects, such as data collection and feature selection, have helped the authors to their successful conclusions. The experiment's researchers have determined the variables that have the most influence on predicting for airplanes, as well as one more variable that could improve forecast accuracy. The accompanying days will see the documentation expanded. to incorporate forecasted prices for all aviation businesses. Airline tickets must be purchased in the best business class, and further analysis of enormous data sets is required. [6] The aim of this work was to find a machine learning model that could better accurately estimate flight costs. After putting various models to the test, it was discovered that the Random Forest algorithm produced an accurate prediction of the outcome. In terms of results, the piece outperforms models, and it aspires to do even better in the future. [7]

## OBJECTIVE

According to our actual notion of the flight price prediction using machine learning is that, we will predict flight fares by comparing today's prices to those of future days. In order to help clients, make informed trip booking decisions, our main objective is to forecast flight costs by comparing today's pricing to any other day. This research-based project will use machine learning as the back end with a substantial quantity of datasets.

## PROPOSED METHODOLOGY

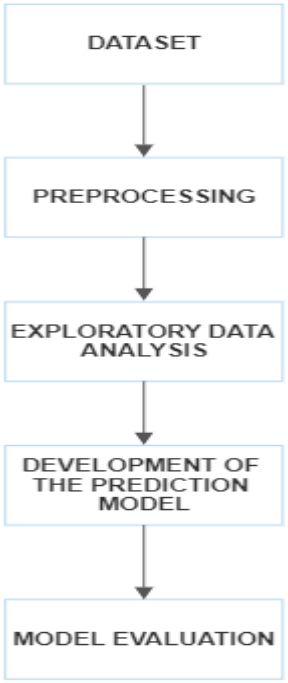


Fig:1 Methodology used for our proposed Flight Price Prediction Using Machine Learning

1.DATASET: Our dataset was acquired from Kaggle, a platform that enables users to locate and post data sets, find and develop models in a data science environment, work together with other data scientists and technical professionals, and participate in competitions to address data science challenges.

2.PREPROCESSING: Preprocessing is used to change the raw data into a useful and practical format. The same information is spread across many attributes in the collection. When the tables are combined directly, several replica fields are created. Furthermore, human mistake, currency conversion errors, etc. could result in inaccurate values being recorded over the airways. Hence, to produce reliable input data so that you may develop a machine learning model, a well-designed records pre-processing approach is crucial.

3.EXPLORATORY DATA ANALYSIS: Data analysis utilizing visual methods is called exploratory data analysis (EDA). With the use of statistical summaries and graphical representations, it is used to identify trends, patterns, or to verify assumptions. In EDA we will do (a) Handling Missing Values (b) Data Visualization (Univariate Analysis, Bivariate Analysis).

4.DEVELOPMENT OF THE PREDICTION OF THE MODEL: To find the most effective algorithm, we will test a variety of them in our project. The support vector machine, the random forest, the decision tree, the KNN, and the linear regression algorithms will all be used. The linear regression algorithm is more effective and significant, according to our analysis of the aforementioned techniques.

5.MODEL EVALUATION: Model evaluation is the act of employing several evaluation measures to comprehend the performance and strengths and weaknesses of a machine learning model. A model's effectiveness must be evaluated in the early stages of research, and model evaluation also contributes to model monitoring.

## EXPERIMENTATION & EVALUATION

1. ***Univariate Analysis:*** In Univariate Analysis we have done the followings:

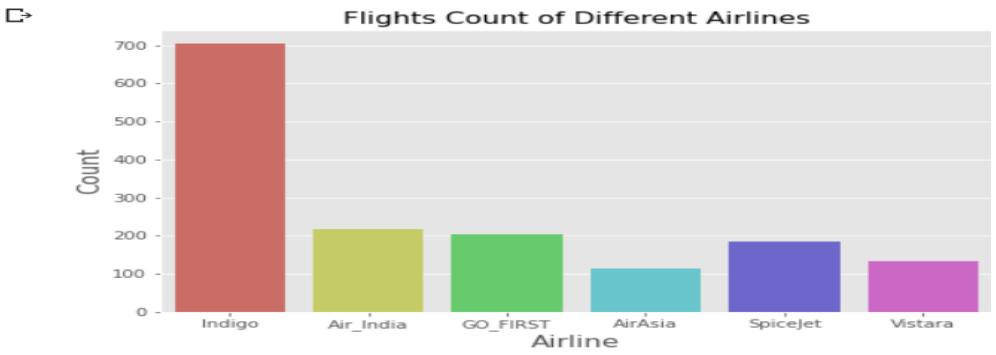


Fig:2 Flight count of different airlines

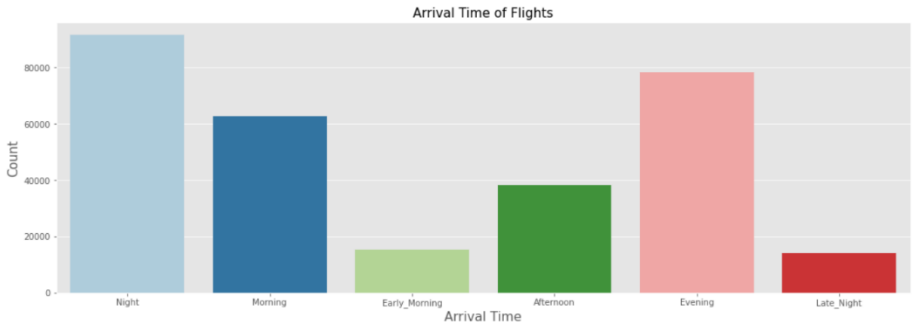


Fig:3 Arrival time of flights

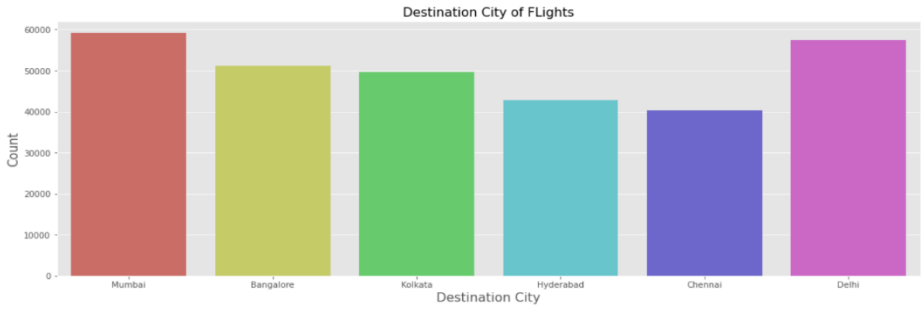


Fig:4 Destination City of Flights

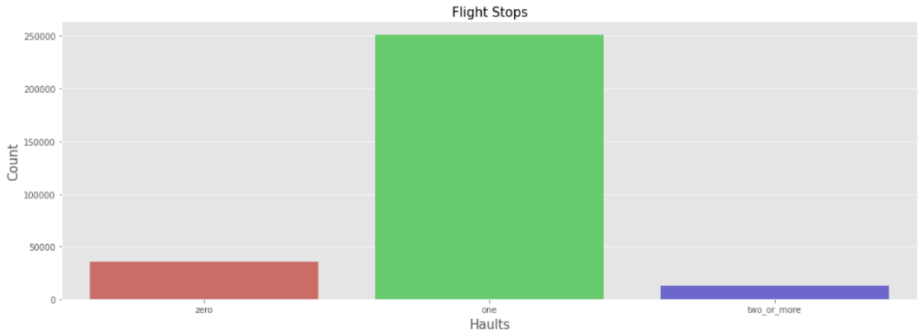


Fig:5 Flight Stops

2. ***Bivariate Analysis:*** In Bivariate Analysis we have done the followings:

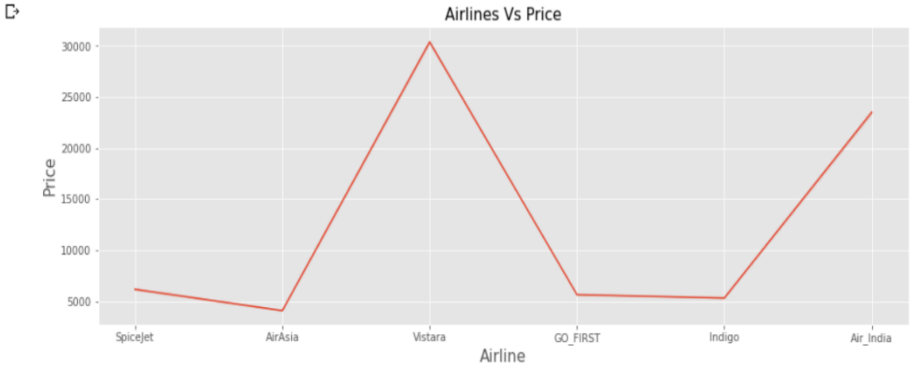


Fig:6 Comparison between Airline and Price

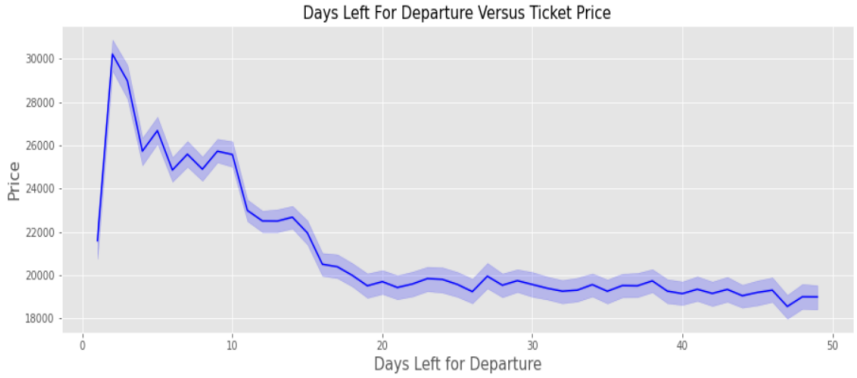


Fig:7 Comparison between Days Left for Departure and Ticket Price


Fig:8 Ticket Price Based on the Departure Time

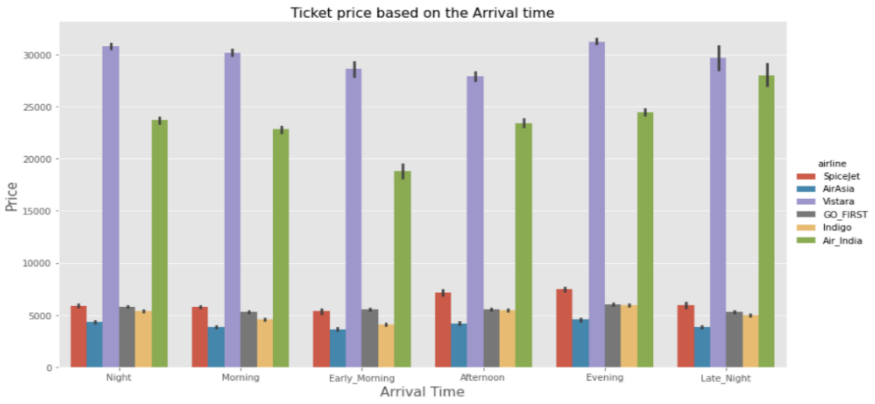


Fig:9 Ticket Price Based on the Arrival Time

# RESULT & DISCUSSION

## 1.MODEL SELECTION & TRAINING:

(1.1)
(a) Storing the Dependent Variables in X and Independent Variable in Y
(b) Splitting the Data into Training set and Testing Set.
(c) Scaling the values to convert the int values to Machine Languages.

(1.2)
(a) Linear Regression**:** One algorithm that corresponds to supervised machine learning is linear regression. On the basis of the data points for the independent variables, it attempts to apply relations that would forecast the outcome of an event. The relation is often a straight line that as closely as possible fits the various data points. A numerical value, is the output.
The mathematical equation of Linear Regression is:

$$y = \beta 0 + \beta\, 1x + \varepsilon \qquad (1)$$

Here
y = Dependent Variable
x = Independent Variable
β0 = Intercept of the line
β1 = Slope of the line
ε = Arbitrary Error

(b) Decision Tree Regressor**:** A form of machine learning model called a decision tree regressor is used to forecast numerical and continuous output values according to input information. It operates by fitting a regression model to each smaller subset of the data after recursively partitioning the data into smaller subgroups based on the values of the input features. The feature that offers the best split of the data is first found via the decision tree regressor algorithm.
The mathematical equation of Decision Tree Regressor is:

$$y = f(x) \qquad (2)$$

Here
y = Predicted Output
x = Input Feature Vector
f = Decision Tree Regressor

(c) Random Forest Regressor**:** A random forest regressor is a sort of machine learning method used for regression problems. A significant number of decision trees are constructed in a random forest regressor using randomly chosen portions of the training data and features. Each tree in the forest is trained to foresee the target variable, and the final prediction is made by averaging all of the trees' predictions. Due to its excellent accuracy, scalability, and simplicity of use, the random forest regressor is a preferred option for many regression situations.

(d) K Neighbors Regressor**:** A machine learning approach used for regression issues is called the K Neighbors Regressor. The K Neighbors Regressor algorithm locates the k data points in the training set that are most comparable to a new data point. The average of the target values of these k neighbours is then used to calculate the anticipated value for the new data point.
The mathematical equation of The K Neighbors Regressor is:

$$\boldsymbol{y} = (\frac{\boldsymbol{1}}{\boldsymbol{k}}) * \sum \boldsymbol{yi} \quad (3)$$

Here

y = Predicted value for a new datapoint
x = Target value for a closest datapoint
k = Number of datapoints for prediction

(1.3)

(a) Mean Squared Error (MSE)**:** How closely a regression line resembles a set of data points is determined by the Mean Squared Error. It is an unpredictable function that relates to the anticipated value of the squared error loss.The average, more particularly the mean, of errors squared from data related to a function is used to determine it.
The mathematical equation of MSE is:

$$\boldsymbol{MSE} = \left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\right) * \boldsymbol{\Sigma i = 1 \; to \; n} \; (\boldsymbol{yi} - \hat{\boldsymbol{y}})^{\boldsymbol{2}} \quad (4)$$

(b) Mean Absolute Error (MAE)**:** One often used metric to assess the effectiveness of a regression model is mean absolute error (MAE). It calculates the average absolute difference between a group of data points' actual and anticipated values.
The mathematical equation of MAE is:

$$\boldsymbol{MAE} = \left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\right) * \sum |\boldsymbol{y} - \hat{y}| \quad (5)$$

(1.4)

The accuracy of each algorithm is shown in the below table

Table:1

| Machine Learning Model | r2score_train | r2score_test |
|---|---|---|
| Linear Regression | 0.9045992290556947 | 0.9266562364171973 |
| Decision Tree Regressor | 0.9899097986789675 | 0.973050949087886 |
| Random Forest Regressor | 0.9984635585302546 | 0.9840909310968284 |
| K Neighbors Regressor | 0.9838506418508982 | 0.962824134518819 |

After analyzing a variety of machine learning models, we can see from the table above that Random Forest Regressor provides the highest accuracy (r2score).

## FUTURE WORK

In the future, this study might be expanded to anticipate ticket prices for the entire airline's flight schedule. Early study shows that machine learning models have the potential to help end users by offering advice on when to purchase tickets so they may make a profit, even though we still need to evaluate these machine learning algorithms on huge airline datasets. Search engine inquiries and social media data from other sources are not taken into account. Future improvements to this method could forecast plane values over the entire flight map. Huge flight data sets necessitate additional testing, and if more data, such as the current availability of seats, could be made available, the projected results would be more precise.

## CONCLUSION

In this study, "flight price prediction" was the topic of a preliminary investigation. Using flight pricing data, we gathered from the Kaggle website, we showed that it is possible to estimate airline rates based on prior ticket data. The results of the experiment show that ML models are a trustworthy tool for predicting airfare prices. Additional crucial variables in determining flight price are data collection, preprocessing, exploratory data analysis, model selection &training. The research revealed which traits had the greatest influence in predicting airfare.

## REFERENCES

[1] K. Tziridis T. Kalampokas G.Papakostas and K. Diamantaras "*Airfare price prediction using machine learning techniques*" *in European Signal Processing Conference (EUSIPCO), DOI: 10.23919/EUSIPCO .2017.8081365L. Li Y. Chen and Z. Li" Yawning detection for monitoring driver fatigue based on two cameras*" Proc. 12th Int. IEEE Intel Conf.. Transp. Syst. pp.1-6Oct.2009–

[2] Supriya Rajankar, Neha sakhrakar and Omprakash rajankar "*Flight fare prediction using machine learning algorithms"* International journal of Engineering Research and Technology (IJERT) June 2019 –

[3] "*A survey on machine learning-based flight pricing prediction*." Supriya Rajankar and Neha Sakharkar –

[4] T. Wang et al., "*A Framework for Airfare Price Prediction: A Machine Learning Approach*," doi: 10.1109/IRI.2019.00041 –

[5] K. Tziridis, T. Kalampokas, G. A. Papakostas and K. I. Diamantaras, "*Airfare prices prediction using machine learning techniques*," doi: 10.23919/EUSIPCO.2017.8081365 –

[6] *"A survey on machine learning-based flight pricing prediction."* Supriya Rajankar and Neha Sakharkar

[7] "*A PROPOSAL FOR INDIAN FLIGHT FARE PREDICTION*" Udhhav Arora, Jaywrat Singh Champawat, and Dr. K. Vijaya –

[8] 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.

[9] R. Ren, Y. Yang, and S. Yuan, "*Prediction of airline ticket price*," University of Stanford, 2014.

[10] T. Wohlfarth, S. Clemenc¸on, F. Roueff, and X. Casellato, "*A data-mining approach to travel price forecasting*," in the 10th international conference on machine learning and applications and workshops, vol. 1, 2011, pp. 84–89

[11] "*Airfare Prices Prediction Using Machine Learning Techniques*", K. Tziridis, Th. Kalampokas, G.A. Papakostas HUMAIN-Lab, 2017 25th European Signal Processing Conference (EUSIPCO)

[12] B. Burger and M. Fuchs, "*Dynamic pricing – A future airline business model*," Journal of Revenue and Pricing Management, vol. 4, no. 1, pp. 39–53, 2005

[13] K. Rama-Murthy, "*Modelling of united states airline fares– using the official airline guide (OAG) and airline origin and destination survey (DB1B)*," Ph.D. dissertation, Virginia Tech, 2006.

[14] B. Derudder and F. Witlox, "*An appraisal of the use of airline data in assessing the world city network: a research notes on data*," Urban Studies, vol. 42, no. 13, pp. 2371–2388, 2005.

[15] 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.