Flight Price Prediction Using Machine Learning

Ankita Panigrahi[1], Rakesh Sharma[2], Sujata Chakravarty[3], Bijay K. Paikaray[4] and Harshvardhan Bhoyar[5]

[123]*Dept. of CSE, Centurion University of Technology and Management, Odisha, India.*
[4]*School of Information & Communication Technology, Medhavi Skills University, Sikkim, India*
[5]*Faculty. of Management Studies, Sri Sri University, Odisha, India*

Abstract

Currently, everyone loves to travel by flights. Going along with the study, the charge of travelling through a plane change now and then which also includes the day and night time. Additionally, it changes with special times of the year or celebration seasons. There are a few unique elements upon which the cost of air transport depends. The salesperson has data regarding each of the variables, however, buyers can get confined information which is not sufficient to foresee the airfare costs. Considering the provisions, for example, time of the day, the number of days remaining and the time of take-off this will provide the perfect time to purchase the plane ticket. The motivation behind this paper is to concentrate on every component that impacts the variations in the costs of this means of transport and how these are connected with the diversity in the airfare. Subsequently, at that point, utilizing this data, construct a framework that can help purchasers when to purchase a ticket. Machine Learning algorithms prove to be the best solution for the above-discussed problems. In this project, there is an implementation of Artificial Neural Network (ANN), LR (Linear Regression), DT (Decision Tree), and RF (Random Forest).

Keywords

Machine Learning Algorithms, airfare, supervised learning, predictions, flight, Linear Regression, Artificial Neural Network, Random Forest.

1. Introduction

A person who already has reserved a ticket for a flight realizes how powerfully the price of the ticket switches [1]. Airline utilizes progressed techniques considered Revenue Management to accomplish a characteristic esteeming technique [2]. The most affordable ticket available changes over a course of time. The expense of the booking may be far and wide. This esteeming technique normally alters the cost according to the different times in a day namely forenoon, evening, or night. Expenses for the flight may similarly alter according to the different seasons in a year like summers, rainy and winters, also during the period of festivals. The buyers would be looking for the cheapest ticket while the outrageous objective of the transporter would be generating more and more revenue. Travelers for the most part attempt to buy the ticket ahead of their departure day. The reason would be their belief that the prices might be the highest when they would make a booking much nearer to the day of their flight but conventionally this isn't verifiable. The buyer might wrap up paying more than they should for a comparable seat. Considering the challenges faced by the travellers for getting an affordable seat, various strategies are utilized which will extract a particular day on which the fare will be the least. For this purpose, Machine Learning comes into the picture. Gini and Groves developed a model using PLSR, to predict the appropriate time to book the seats [3]. They extracted their data from well-known booking websites from 22/02/2011 to 23/06/2011.

ACI'22: Workshop on Advances in Computation Intelligence, its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United States
EMAIL: 190301250002@cutm.ac.in (A. 1); 190301120079@cutm.ac.in (A.2); chakravartys69@gmail.com (A.3); bijaypaikaray87@gmail.com (A. 4); harshvardhan.b@srisriuniversity.edu.in (A. 5)
ORCID: 0000-0001-5843-0335 (A. 4)


Using the Linear Quantile Blended Regression methodology, Janssen [4] developed an assumption model for the route of San Francisco to New York with already available data on flight fares for each day provided by www.infare.com. The two important features were the day count from departure and which day of the week it is, whether it's weekday or weekend. This model was capable enough to predict the expense for the flight for the days that were nowhere close to the day of departure but the results were not satisfying if it would be close to the date of journey. A ticket-purchasing time incremental model depending upon marked point processors and information extracting systems and computable investigation strategy was suggested by Wohlfarth [5]. The proposed system changes the heterogeneous value arrangement information to added value arrangement system. For choosing the best synchronizing group and later comparison of advancement model a tree-based order calculation has been used. Papadakis [6] anticipated whether there would be a fall in the airfare later on by addressing the issue as a classification task using Logistic Regression, Linear SVM and Ripple Down Rule Learner models. Ren, Yang, and Yuan [7] worked on Linear Regression, Naïve Bayes, SoftMax Regression, and SVM models in predicting the prices.

2. Data Collection

The assortment of data is the very first step in machine learning projects. There are various sources of data available on numerous websites that are deployed to construct the models. These sites supply a huge variety of data regarding different airlines, routes, times, and tolls. In this part, data gathered from the various available sources are studied. For the execution of this, information is brought from a site called Kaggle. For the assortment of the data and to execute the model's Python is utilized [8-15]. The dataset collected contains information about different airlines in India. It consists of various factors which affect the price of a flight ticket including the price for a particular flight. It contains 10683 rows of data. The features present in the dataset are the name of companies, Date of travelling, Origin, terminus, path of travelling, Time of Departure, Time of Arrival, Travelling Hours, Total Stoppage, Additional Info, and Price.

3. Cleaning and Preparing of Data

Cleaning and preparing data are a very important step in machine learning. The data collected can't be used raw as it may contain certain parameters which would be of no use and also certain data can't be used the way it would be present in the dataset. So, before proceeding to the actual work, the data needs to be filtered and it should be absolutely clean. For achieving this, all the duplicate and null values are removed from the dataset and specific data is converted to a usable format.

4. Machine Learning Techniques

Various conventional machine learning algorithms are used for creating a model for flight fare prediction which is ANN, LR, DT, and RF. These loads of machine learning techniques are executed using the sci-kit-learn library available in python. For assessing the exhibition of these algorithms, definite boundaries are thought of. These are mentioned as follows: MAPE (Mean Absolute Percentage Error) and RMSE (Root Mean Square Error).

4.1 RMSE

RMSE is a tool that helps in determining how accurately the model is making the predictions. It calculates how much error the model creates while making these predictions. It measures the standard of predictions. Mathematically, it is defined as the square root of the average of the squares of all the errors. Error is defined as the difference between the actual and predicted value. Less the RMSE, the better the performance of the model is. Usually, an RMSE score of less than 1 is considered the best.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \quad (1)$$

$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are predicted values

$y_1, y_2, \ldots, y_n$ are observed values

n is the number of observations

4.2 MAPE

Mean Absolute Percentage Error is most often used in regression problems. It is most popular in calculating errors in forecasting. It gives an idea about how much accurately the model is evaluating the predictions. Statistically, it is the mean or average of the absolute percentage errors of forecasts. Error is characterized as the contrast between the actual and predicted value. Less the MAPE, the better the exhibition of the model is. Typically, a MAPE score of below 1 is viewed as awesome.

$$\boldsymbol{MAPE = \frac{100}{N} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|} \quad (2)$$

Here,

A_t is the actual-value

F_t is the forecasted-value

5. Machine Learning Algorithms Used

5.1 Artificial Neural Network (ANN)

An artificial Neural Network is simply a Neural Network that resembles the biological Neural Network present in the human brain. It is designed in a way such that it would function the same way a human brain function. It is the collection of millions and millions of artificial neurons. These artificial neurons are the building blocks of the ANN model. Artificial Neuron consists of Inputs and their corresponding weights. An activation function is chosen which takes these inputs multiplies them to their corresponding weights and produces the output. Every Artificial Neural Network must have three layers: the input layer which takes the input, the hidden layer where all the computations take place, and the output layers which produce the output.

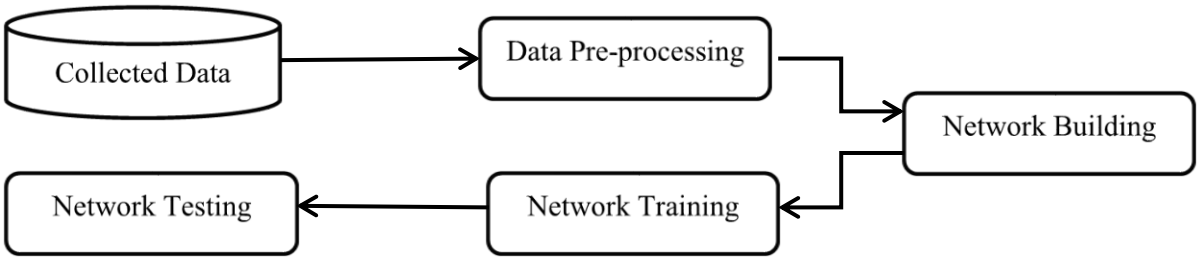

Figure 1. Flow Diagram of ANN Model

In the case of hidden layers, we have used Relu as Activation function with 20 and 10 for weights whereas Linear Activation Function with weight 1 is used in case of the final output. Here, adam optimizer is used.

$$Z_i = \left(\sum_{k=1}^{N_{j-1}} X_k^{j-1} W_{k,i} - b_k\right) \quad (3)$$

$$f(Z_i) = \frac{1}{1 + e^{-Z_i}} \tag{4}$$

5.2 Linear Regression

Linear Regression is an algorithm in machine learning. It works by finding the relationship between single or multiple input variables and the output variable. These relationships are built with linear predictor functions. The graph of a linear regression model is linear justifying its name.

$$y(predicted) = b0 + b1 * x \tag{5}$$

Here,

y is dependent variable,
x is independent variable,
b0 is constant,
b1 is slope.

5.3 Decision Tree

This model is a member of a supervised learning family. It can fit well in both classification and regression problems. As its name says, it is structured like a tree containing the decision nodes and leaf nodes. Decision nodes have multiple branches for decision making where leaf nodes represent the outcomes of these decisions which is further not divided into any branches.

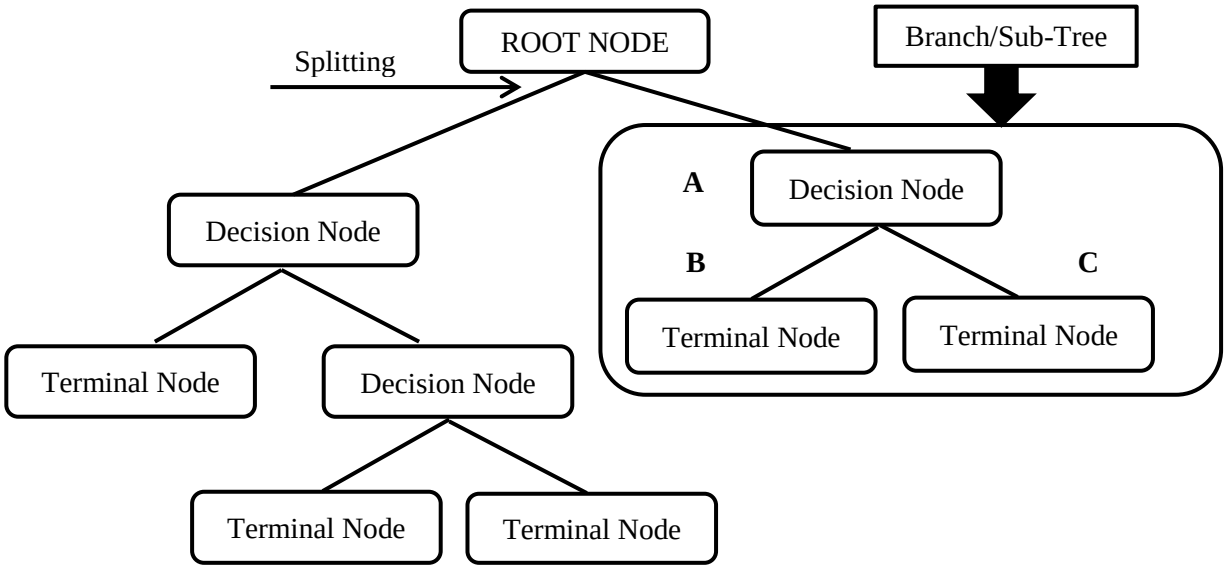

Figure 2. Decision Tree Process

If we write mathematically,
Entropy having 1 attribute:

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \tag{6}$$

Entropy having multiple attributes:

$$E(T, X) = \sum_{c \in X} P(c) E(c) \tag{7}$$

5.4 Random Forest

Like the Decision Tree, RF is also a supervised learning technique. Random forest works with multiple decision trees. Here, the trees are operated as an ensemble. Every tree present in a random forest divides a class prediction and the class having the most votes comes out as models' prediction.

6. Algorithms Evaluation

On comparing the Root Mean Square Errors of the pre-processed data when applied on proposed algorithms it is specified that Artificial Neural Network gives 0.008410 followed by Random Forest giving 0.006240 then Linear regression with 0.006109 closely baking up by Decision Tree with the least error of 0.006101 which shows the Decision Tree model works more precisely than others when applied on the given data. The value given in Table 1 is graphically represented in Figure3.

Table 1

Different ML Models RMSE Errors

ML Algorithms	RMSE
Artificial Neural Network	0.008410309713082834
Linear Regression	0.006109087698177261
Decision Tree	0.0061019746207730645
Random Forest	0.0062402313794453

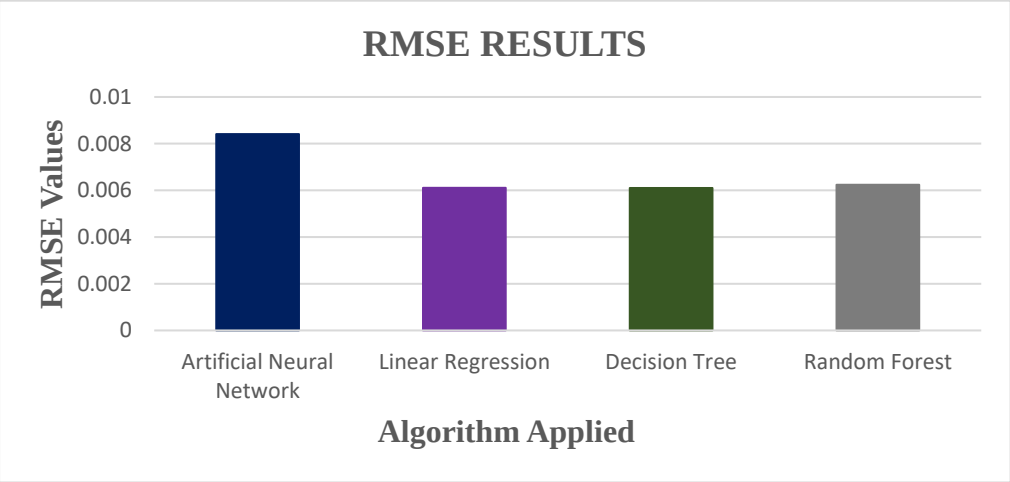

Figure 3. Result Analysis of RMSE for all applied Models

If we go through the Mean Absolute Percentage Error the results show that again the Decision Tree got the least Error when compared with all other models. The value given in Table 2 is graphically represented in Figure 4.

Table 2

Different ML Models MAPE Errors

ML Algorithms	MAPE
Artificial Neural Network	0.6831663296497983
Linear Regression	0.5202171579180117
Decision Tree	0.5202012748283965
Random Forest	0.5291447939343533

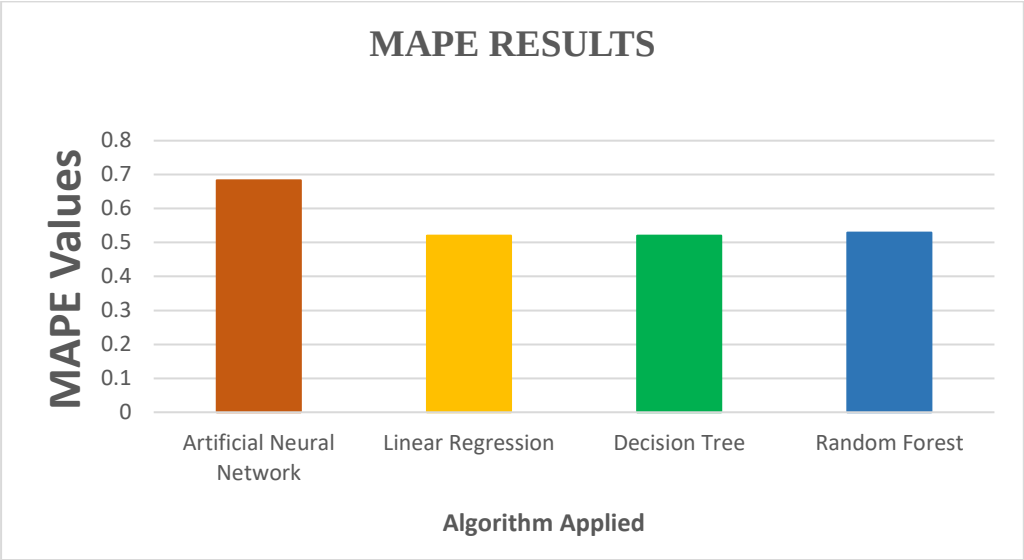

Figure 4. Result Analysis of MAPE for all applied Models

7. Conclusion

We learn that ML models can be used to predict prices based on earlier data more correctly. The presented paper reflects the dynamic change in the cost of flight tickets from which we get the information about the increase or decrease in the price as per the days, weekends, and the time of the day. With the Ml algorithm applied on various datasets, better results can be obtained for prediction. The error values that we got for Artificial Neural Network are comparatively high but for obtaining lesser values we can use evolutionary algorithms of ANN like genetic algorithms in the future.

8. References

[1] Rajankar, Supriya, and Neha Sakharkar. "A Survey on Flight Pricing Prediction using Machine Learning." Internatıonal Journal Of Engıneerıng Research & Technology (Ijert) 8.6 (2019): 1281-1284.

[2] Smith, Barry C., John F. Leimkuhler, and Ross M. Darrow. "Yield management at American airlines." interfaces 22.1 (1992): 8-31.

[3] Groves, William, and Maria Gini. "An agent for optimizing airline ticket purchasing." Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. 2013.

[4] Janssen, Tim, et al. "A linear quantile mixed regression model for prediction of airline ticket prices." Radboud University (2014).

[5] Wohlfarth, Till, et al. "A data-mining approach to travel price forecasting." 2011 10th International Conference on Machine Learning and Applications and Workshops. Vol. 1. IEEE, 2011.

[6] Papadakis, Manolis. "Predicting Airfare Prices." (2014).

[7] Ren, Ruixuan, Yunzhe Yang, and Shenli Yuan. "Prediction of airline ticket price." University of Stanford (2014).

[8] Tziridis, Konstantinos, et al. "Airfare prices prediction using machine learning techniques." 2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017.

[9] Boruah, Abhijit, et al. "A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2019. 191-203.

[10] S. Chakravarty, B. K. Paikaray, R. Mishra and S. Dash, "Hyperspectral Image Classification using Spectral Angle Mapper," 2021 IEEE International Women in Engineering (WIE) Conference on

Electrical and Computer Engineering (WIECON-ECE), 2021, pp. 87-90, doi: 10.1109/WIECON-ECE54711.2021.9829585.

[11] Wang, Tianyi, et al. "A framework for airfare price prediction: A machine learning approach." 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2019.

[12] Abdella, Juhar Ahmed, et al. "Airline ticket price and demand prediction: A survey." Journal of King Saud University-Computer and Information Sciences 33.4 (2021): 375-391.

[13] Zhao-Jun, Gu, Wang Shuang, and Zhao Yi. "Flight ticket fare prediction model based on time-serial." Journal of Civil Aviation University of China 31.2 (2013): 80.

[14] Huang, Tenghui, Chih-Chien Chen, and Zvi Schwartz. "Do I book at exactly the right time? Airfare forecast accuracy across three price-prediction platforms." Journal of Revenue and Pricing Management 18.4 (2019): 281-290.

[15] S. Chakravarty, P. Mohapatra, P. K. Dash, (2016), Evolutionary Extreme Learning Machine for Energy Price Forecasting, International Journal of Knowledge-Based and Intelligent Engineering Systems, 20, 75-96

[16] https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh/

[17] https://github.com/rishabdhar12/Flight-Price-Prediction/tree/main/Dataset