International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

# Flight Ticket Price Predictor Using Machine Learning

[1]T.Sravya Sri, [2]S.sindhuja, [3]R venu kumar, [4]S.yadagiri, [5]N.Srikanth,
[1,2,3,4,]U.G.Scholor, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.
[5]Research Guide, Department of ECE, Sri Indu College Of Engineering & Technology, Ibrahimpatnam, Hyderabad.

## ABSTRACT

This study primarily aims to assess two different methods for estimating the cost of train tickets: a Decision Tree & a Novel Random Forest. Procedures and Materials: Decision Tree (DT) and Random Forest (RF) are the two categories that make up this research. The six iterations for each group were determined using ClinCal with the following parameters: alpha = 0.05, beta = 0.2, Gpower = 0.8, and a 95% confidence interval. We used Kaggle's 215909-item train ticket price prediction dataset. In comparison to Decision Tree's accuracy rate of 78.18%, Random Forest's algorithm achieves a much higher rate of 88.01%. The Decision Tree as well as the Novel Random Forest approach differ considerably from each other ($p=0.001$), according to the findings for the t-test on independent samples ($p<0.05$). Findings: When comparing the two models, Random Forest outperforms Decision Tree when it comes to forecasting the price of train tickets.

Machine learning, decision trees, insurance, public transportation, and novel random forests are some of the terms associated with train ticket prices.

## INTRODUCTION

Modern revenue management strategies used by railroad companies include train ticket pricing. Its goal is to increase profits by quickly calculating passenger demand and adjusting ticket prices accordingly [1]. In addition, modern intelligent public transportation networks cannot be built without consideration of train ticket price and insurance. Train ticket price is also crucial to the creation of an intelligent public transportation system that can control congestion, manage mobility, and lower peak loads [2]. Reducing peak loads and controlling congestion while keeping costs low is the goal of mobility management. At the same time, travelers are deciding when it's cheapest to buy a certain ticket and whether or not to add insurance to it. Predicting ticket costs, for example in railway or airline situations, becomes particularly relevant in this context [3]. The most popular approach to fixing this problem is to use past prices as a predictor variable along with the chosen train, plane, or bus departure details, such as the insurance price (such as the amount of days before the departure and the departure weekday) [4]. It also seems like available Internet data from things like social media and search engine inquiries is still not being used to improve the precision of predictions for dynamic insurance and ticket pricing. A

International journal of imaging science and engineering
ISSN: 1934--9955
Vol-20 Issue-01 Mar

lot of people utilized apps like IRCTC, Make My Trip, and TrainMan to find out how much train tickets were and to buy them [5].

There have been 103 publications in Google Scholar and 154 in IEEE Xplore throughout the last five years. This is correct for the dynamic pricing of train tickets it would get text messages for mobiles [7], In order to address social, economic, and environmental concerns on a national and even international scale, this book contributes to the development of urban rail transportation that is fast, dependable, and energy efficient [8]. A critical next step is to put a stop to ticket speculation and boost the efficiency of insurance ticket sales [5] that can reliably forecast ticket prices according to variables such travel distance, season, day of the week, and demand [3].
The research study found that using a decision tree technique to machine learning for classifier calculations resulted in less accurate predictions of rail ticket pricing. Proficiency in using a machine learning classifier to forecast the cost of locomotive tickets The suggested study is more accurate because to the novel random forest approach. In order to provide customers the greatest and most accurate forecasts quickly, this project aims to develop a new system for public transportation that predicts railway ticket costs. The objective for ticket price prediction was to help customers out by showing them the prices of train tickets without them having to sift through all the applications. from the limited but growing literature on intelligent public transportation systems, which provides a reference design model for bettering public transportation services. Public transportation operators, especially those in developing countries, may benefit from the models shown here as they give ideas and strategies on how to enhance and ensure the safety and comfort of riders.

## MATERIALS AND METHODS

Data analytics at Saveetha School of Engineering's Computer Science Department was responsible for collecting and analyzing the dataset. The research sample was split into two categories: Decision Tree and Novel Random Forest. We took six samples from each group. Statistical parameters used in this study are G-power 0.8, alpha = 0.05, beta = 0.2, and a confidence interval of 95%. The Kaggle Train Ticket price prediction dataset, including 21,5909 records [10], was used for this investigation.

While 80% of the dataset was used for RF training, only around 20% was utilized for testing. When training, the RF set uses back propagation, as opposed to the DT model's hyper plane-based data segmentation. Now that dataset fitting is complete, the DT and RF models may be trained. We tested the two models' performance on a 6-item dataset in Python 2.7.

### Technique for Randomization Decision Tree

International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

After reducing the size of the input data set, Random Forest takes an average of the results from many decision trees to boost the accuracy of the dataset's predictions. The random forest obtains its final output by combining data from all the trees and applying the forecasts that are utilized the most. Due to the reduced likelihood of over fitting in a wider forest, accuracy is enhanced. You can see the Random Forest Algorithm's pseudo code in Table 1. The second thing to do is make sure that all of the variables are defined properly; for instance, we might call X a train and Y a forecast. At last, activation functions—also called the sigmoid function—determine the output. We make use of it to ensure that our forecasts are as precise as possible.

**Method for Analyzing Decision Trees**

Issues with both classification and regression may be addressed concurrently when this method is used. In an effort to find a solution, the decision tree approach uses a graphical representation of a tree. The nodes within the tree stand for the attributes, whilst the nodes outside the tree represent the labels for the classes. Table 2 shows the decision tree algorithm's pseudo code. It is therefore possible to use the result to create the most precise predictions.

**Statistical Analysis**

The decision tree and new random forest algorithms in SPSS Version 23 were used to statistically compare the permissible and forbidden categories [11]. Using throughput analysis, we determined the independent test, standard deviation, average widow size without using any dependent parameters.

**RESULTS**

The accuracy of the two approaches, Random Forest (88.01%) and Decision Tree (78.88%), are shown in Table 3. Both methods use a sample size of 6. In comparison to the decision tree method, the suggested Random Forest approach outperforms it.

The statistical results for the two groups' means, standard deviations, and standard errors are shown in Table 4. An average accuracy of the RFM is 88.01%, whereas the accuracy of decision trees is 78.88%.

As shown in Table 5, the accuracy and loss of the dataset were examined using using a 95% confidence interval in an independent sample t-test for group difference. A two-tailed result of $p=0.001$ ($p<0.05$) indicates that the Decision Tree and the Novel Random Forest method vary significantly from one another.

International journal of imaging science and engineering
ISSN: 1934--9955
Vol-20 Issue-01 Mar

The graph's two sets, representing the Random Forest and decision tree methods, are shown in Figure 1. Decision trees' mean accuracy is lower and their error range is ±2 SD when contrasted with random forest methods.

## DISCUSSION

With the use of Decision Tree (78.88%) and Novel Random Forest (88.01%), this research has developed a method for reliably predicting ticket prices. Random Forest seems to be more accurate than Decision Tree. In terms of accurately predicting ticket prices, Random Forest seems to be more effective than Decision Tree. Given that both groups' p-values are less than 0.05 ($p<0.05$) and 0.001, it may be inferred that they both have a significant statistical influence.

This research has the support of several published investigations. Carrying the duties delegated by [12]. Preliminary algorithms, artificial features, classifiers, deep learning, and associated technologies will all be a part of this research on train ticket pricing [13]. After this, we will review further studies that thoroughly analyzed the cost of rail tickets. Lastly, we go over the databases utilized for rail tickets and the guidelines for overall evaluation. stated as [14]. This research explains how to build a public transit train ticket system using OpenCV, a Python package, and deep learning. A multi-view tickets database has been requested by several individuals, and ML methods such as random forest have been praised in various articles for their predictive capacities [15]. The authors concede that acquiring and processing large amounts of data could be time-consuming and expensive for proper algorithm training.

Unfortunately, reliable data is hard to come by, and there are a lot of variables (like insurance) that affect ticket prices, making it difficult to forecast how much train tickets will cost. In order to provide more user-friendly and trustworthy predictions, scientists should keep working to enhance the model.

## CONCLUSION

This research utilizes the Decision Tree (78.88) and Random Forest (88.01) to forecast ticket prices. It would seem that Random Forest is better at predicting ticket prices than Decision Tree.

## DECLARATION

### Possibility of Forgery

This study does not include any possible prejudice.

### Authors' Contributed Works

ISSN: 1934--9955
Vol-20 Issue-01 Mar

It was author MVR's job to gather data, write the paper, and analyze it. The manuscript's critical review, data validation, and general conceptualization were all SMK's doing.

**Expressing Appreciation**

The faculty and staff of Saveetha University's Saveetha Institute of Medical and Technical Science and Saveetha School of Engineering were very helpful to me during this endeavor, and I am very grateful to them.

**Funding**

Thanks to the financial support of the following organizations, we were able to complete the study.
II. Qbec Infosol, based in Chennai.
Second, Saveetha University
The Saveetha Institute for Health and Technology, number three.
Engineering at Saveetha University: 4.

International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

## REFERENCES

[1] A. Serusi, *The Impact of Ride-Hailing Services on Real Estate Prices in Urban Districts*. 2022.
[2] S. Sharma, A. K. Jain, and S. Devendra, *EXCELLENCE IN METRO OPERATIONS AND MANAGEMENT: BEST PRACTICES WORLD OVER*. PHI Learning Pvt. Ltd., 2022.
[3] W. Zhou and X. Han, "Integrated optimization of train ticket allocation and OD-shared ticket sales strategy under stochastic demand," *Transportation Letters*. pp. 1–10, 2023. doi: 10.1080/19427867.2023.2174662.
[4] Y. Hou, T. Liu, Z. Zhao, and Y. Wen, "Estimating the Cultural Value of Wild Animals in the Qinling Mountains, China: A Choice Experiment," *Animals (Basel)*, vol. 10, no. 12, Dec. 2020, doi: 10.3390/ani10122422.
[5] F. T. Guides, *Fodor's London 2020*. Fodor's Travel, 2020.
[6] Jin, Jin, and You, "Do Discounts in Ticket Prices Induce Sustainable Profit to Performing Arts Suppliers?," *Sustainability*, vol. 11, no. 14. p. 3829, 2019. doi: 10.3390/su11143829.
[7] C. Ananthakumar, D. Edison Selvaraj, and E. Krishnamoorthy, "Metro Train Safety and Ticket Checking System," *International Journal of Science and Engineering Applications*, vol. 8, no. 3. pp. 90–92, 2019. doi: 10.7753/ijsea0803.1001.
[8] M. Setiyo *et al.*, *BIS-HSS 2020: Proceedings of the 2nd Borobudur International Symposium on Humanities and Social Sciences, BIS-HSS 2020, 18 November 2020, Magelang, Central Java, Indonesia*. European Alliance for Innovation, 2021.
[9] "Flight Ticket Price Prediction using Machine Learning," *Journal of Xidian University*, vol. 14, no. 6. 2020. doi: 10.37896/jxu14.6/289.
[10] "Train Ticket Price Prediction project," Jul. 05, 2022. https://kaggle.com/code/umairaslam/train-ticket-price-prediction-project (accessed Apr. 18, 2023).
[11] R. P. S. Kaurav, D. Gursoy, and N. Chowdhary, *An SPSS Guide for Tourism, Hospitality and Events Researchers*. Routledge, 2020.
[12] E. Santana, S. Mastelini, and S. Jr., "Deep Regressor Stacking for Air Ticket Prices Prediction," *Anais do Simpósio Brasileiro de Sistemas de Informação (SBSI)*. 2017. doi: 10.5753/sbsi.2017.6022.

International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

## TABLES AND FIGURES

### The Random Forest Method's Implicit Program (Table 1)

| **I. Input dataset records** |
|---|
| 1. Bring in all of the necessary packages. |
| Step two: following the extraction feature, transform the picture files into numerical numbers. |
| Thirdly, make sure that the four variables—X_train, y_train, X_test, and y_test—are available. in the dataset. |
| Finally, send the training & testing variables to the train_test_split() method. |
| 5. To use RF training to divide the data, use test_size and random_state as arguments. |
| 6. Loading the sklearn library with life. |
| 7. Make a prediction based on the testing data using Classifier. |
| 8. Determine the model's accuracy. |
| **OUTPUT**<br>Accuracy |

International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

**Section 2. DTA Pseudo code.**

| **Decision tree algorithm** |
|---|
| 1. The required characteristics are defined by x and y. |
| secondly, dt is equal to decision tree classifier with random state x. |
| A third step is to utilize the dt.fit function with the parameters X_train and y_train, with y_pred_dt equal to dt.predict(X_test). |
| 4. The value of current accuracy is calculated by rounding the accuracy score of Y_pred_dt and Y_test by 100 and then adding 2. |
| 6. if the current accuracy is greater than the maximum accuracy: |
| 6. The present accuracy is equal to the maximum accuracy. |
| 7. x for best_x |
| Decision tree classifier using best_x as the random state |
| fitting dt provided Two variables, X_train and Y_train |
| This brings us to our eleventh condition: Y_pred_dt=dt.predict(X_test). |

**Table 3. Accuracy of Predictions for Train Ticket Prices**.

| **Number of observations** | **How accurate is the random forest method as a percentage?** | **Proportional accuracy of decision tree algorithms** |
|---|---|---|
| 1 | 88.01 | 78.88 |
| 2 | 87.50 | 78 .92 |
| 3 | 86.70 | 75.88 |

International journal of imaging science and engineering
ISSN: 1934--9955
Vol-20 Issue-01 Mar

| | | |
|---|---|---|
| 4 | 85.70 | 73.76 |
| 5 | 81.08 | 77.92 |
| 6 | 80.09 | 72.63 |

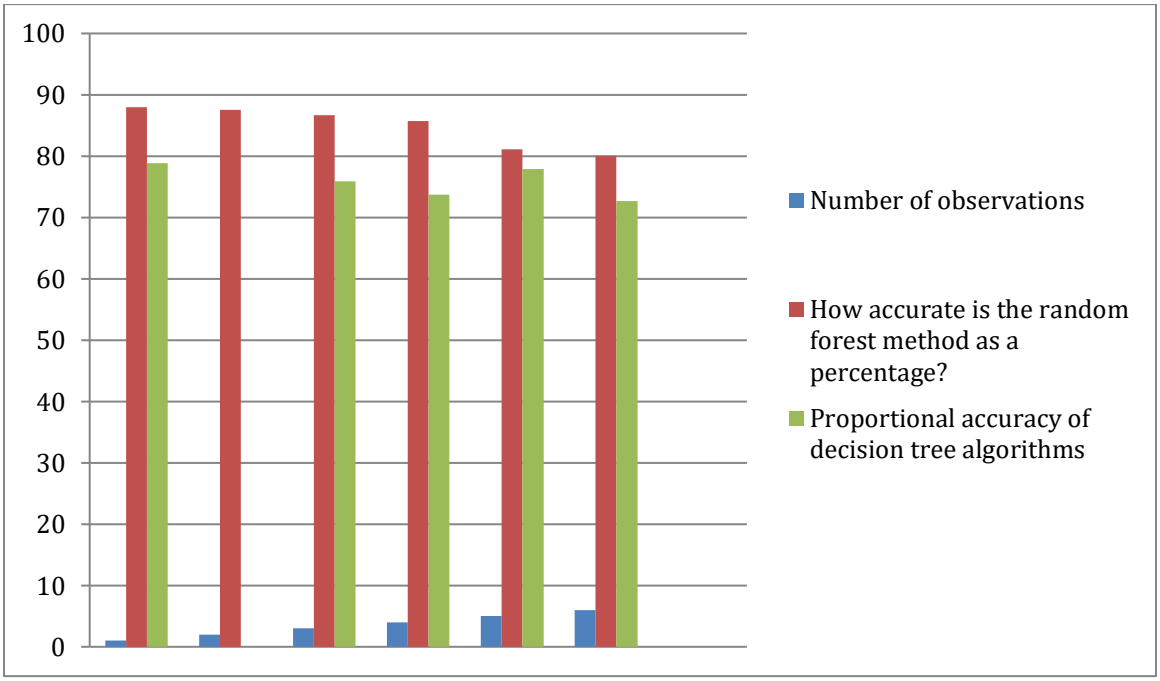


**fig. 1 Accuracy of Predictions for Train Ticket Prices**.

**Table 4.** Analysis of variance, standard error, and mean for a group

| **Groups** | **N** | **Mean** | **Average Distinction** | **Typical Error Mean** |
|---|---|---|---|---|
| Accuracy Random Forest | 6 | 84.8467 | 3.40657 | 1.39073 |
| Decision Tree | 6 | 76.3317 | 2.69207 | 1.09903 |

**Table 5.** A two-tailed result of p=0.001 (p<0.05) indicates that the Novel Random Forest method differs significantly from Decision Tree, according to an independent T sample test.

International journal of imaging science and engineering

ISSN: 1934--9955
Vol-20 Issue-01 Mar

| | | Using Levine's Test to Verify Equality of Variance | | Comparison of Means via T-test | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Distinct 95% CI of the disparity | Different 95% CIs for the difference |
| | | F | Sig. | t | df | Typed twice | Comparative Means | Standard Deviations in Errors | lower | upper |
| Accuracy | Assumption of Equal Variances | .688 | .426 | 4.804 | 10 | .001 | 8.51500 | 1.77257 | 4.56548 | 12.4645 |
| | No assumption of equal variances made | | | 4.804 | 9.493 | .001 | 8.51500 | 1.77257 | 4.53669 | 12.4933 |

International journal of imaging science and engineering
ISSN: 1934--9955
Vol-20 Issue-01 Mar

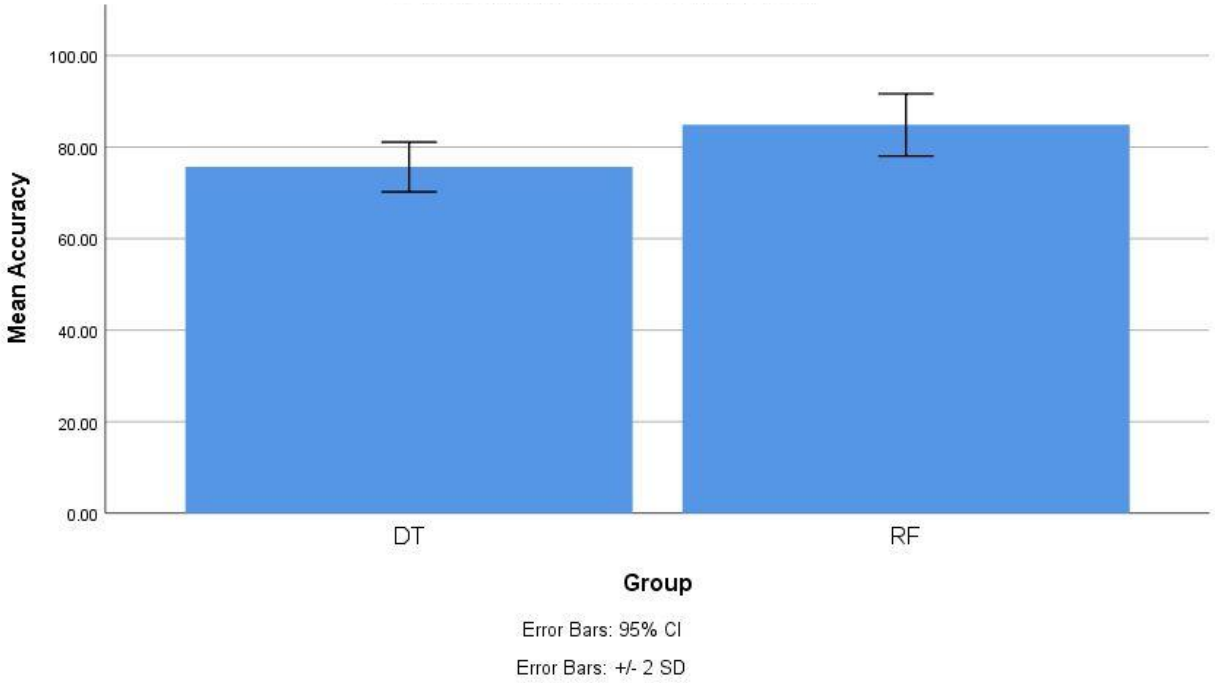


**Fig. 2.** The graph's two sets of comparisons reveal that Decision Tree techniques outperform Random Forest in terms of Mean Accuracy. One axis  Decision vs. Random Forest the y-axis, and the error bar is ±2 standard deviations.