PREDICTING OPTIMAL TICKET PRICING: A MACHINE LEARNING FRAMEWORK FOR AIRLINE FARE ANALYSIS

Mrs.S.Poovidha Assistant Professor(CSE) poovidhasubramaniyan@gmail.com
Mrs.Suchitra Devi A Assistant Professor(CSE) a.suchitradevi27@gmail.com
Dr.T.John Peter Professor (CSE) cshodsait@gmail.com

Annotation

With the increasing popularity of air travel in India and the ability to purchase numerous flight tickets online, passengers are trying to understand how these airline companies determine the pricing of their tickets. There are several tactics that enable you to take action at the appropriate time. Consumers seek the lowest ticket price, but airlines aim to maintain as high an income as possible in order to maximize their earnings. To increase income, airlines use a range of computational techniques, such as pricing discrimination and demand forecasting. This is for the customer who estimates the cost of the flight while purchasing a ticket. From the standpoint of the client, the main challenge is determining the best bargain or the best time to Buying tickets is the most challenging part. The majority of the methods rely on prediction models, machine learning (ML), and sophisticated computational intelligence. This study highlights the elements and offers guidance for creating an aviation fare prediction model based on machine learning.

Keywords: Hyper parameter tweaking, Random Forest, Machine Learning, and flight fare

INTRODUCTION

One of the main motivations is travel, as Everyone aspires to have new experiences in life.

India, with its variety, provides one of the world's most intriguing cities, an abundance of amazing alternatives, and one of the world's most diversified countries in terms of religion and ethnicity. India is one of the most promising nations in the world because of its diverse culture and geography. Regarding the price of plane tickets, these trends also give rise to worries. Making an informed decision may be challenging when comparing the price of a plane ticket from yesterday to today. Travellers who wish to view a new place in India, one should be informed of the cost of tickets in order to get the most affordable and dependable pricing. That satisfies their needs. The idea of predicting airline tickets was born out of this gap, which aims to facilitate passengers' ability to purchase tickets that satisfy their needs. The project's data collection includes 13-column records totaling 10683 include all domestic and international flights that took place in India during the year in 2019. To predict the cost of a plane ticket depending on the data columns' attributes, we have evaluated this data collection using methods for machine learning in this study. Based on these columns, we will try to remove null values (error input) so that they have no effect on the outcome of our analysis. To determine the number of persons in each group, we will begin by looking at the Duration column. Before eliminating the Nan values, In search of null values. The process of implementation continues with exploratory data analysis, which is the following phase. critical to meeting our prediction. We plan to incorporate specific features that can be utilized to characterize the results of our machine learning model. Among the most important parts of EDA is handling categorical data. Ordinal data, which employs Label and hot encoders in one (with order) and as well as nominal data (out of order), are two categories for categorical data. We will utilize One Hot Encoder for the Airline column due to the nominal categorical data. The Price and Source properties will be compared when the same code is used in the Source column. Our machine learning procedure uses test and training datasets were previously divided. This strategy will be used during the exploratory data analysis stage of the implementation process, which is critical to meeting our prediction. We plan to incorporate specific features that can be utilized to characterize the results of our machine learning model. Among the most important parts of EDA is handling categorical data. Ordinal data, which employs Label and hot encoders in one (with order) as well as nominal data (out of order), are two categories for categorical data. We will utilize One Hot Encoder for the Airline column due to the nominal categorical data. Applying the same code to the Source column will result in a comparison of the Price and Source attributes. Our machine learning uses the training and test datasets. Methodology. were previously divided. This tactic will be applied to stop the leakage of data. After being imported, the data is preprocessed. We will forecast the price in this study. In order to ascertain which characteristic has the strongest correlation with the goal variable (price), we will conduct a feature selection process. For this kind of study, we'll use

Random Forest and scikit-learn to segment the data. Hyperparameter adjustment, which comprises Grid Search CV and Randomized Search CV, can raise this score. This study examines the top three designs and approaches to learning. With our project, we are able to overcome this difficulty. Section 1 of the additional documentation provides an explanation of the literature study, and Section 2 provides an explanation of the suggested system architecture. We will walk through the architectural process and algorithm.

LITERATURE SURVEY

Tiyani Wang [1] suggested forecasting costs based on price at the marketing strategy level. the datasets DB1B and T-100, as well as economic data. The main elements of the suggested framework are briefly summarized. EVERY DATASET IS eliminated to eliminate any erroneous sample data during the data preparation step, and they are then modified and combined according to the market segment. In order to define a market segment, the feature extraction module gathers and creates handcrafted attributes. The study component's output, a hybrid web application development platform with two APIs and a web application developed using React native, was proposed by P.H.K.

Tissera [2]. While Python Flask is used to write one API, Node.js is used to write the other Kalampokas[3] and K. Tziridis[4] suggested the intricate strategies and methods that airline companies employ to determine dynamic airfare pricing. These Strategies consider a range of

financial, marketing, commercial, and social aspects that have an impact on the flight's final cost. Because airfares vary, it might be difficult for a customer to get the greatest value on an airline ticket. frequently due to the extremely complex pricing processes used by airlines. G.A. Papakostas [5] suggested a number of methods that, by estimating the cost of the flight, might indicate when it is best for customer to buy an airline ticket. Most of these tactics rely on sophisticated prediction models created in the computational intelligence research (ML) field of machine learning.

A model of a linear quantile hybrid regressor developed by Janssen [6] has good predictive power for airline ticket prices up to several days in advance of departure. In their investigation, Ren, Yang, and Yuan [7] discovered that the LR (77.06% acc.), NB (73.06%

acc.), SR (76.84% acc.), and SVM (80.6% acc. for two bins) models performed well in forecasting the price of airline tickets.

PROPOSED SYSTEM

The Random Forest Algorithm was applied in the suggested system. It may be used for regression and classification issues with machine learning. It is supported by the ensemble learning concept. Large datasets with a high dimensionality can be handled by it. We can apply a strategy to find the cheapest airline ticket even with large data sets. Extra-trees regressors are used to generate a para estimator, which fits several decision trees that are randomly picked on different subsamples of the data set. Aggregating is then used to boost predicted accuracy and control over-fitting. We used a scatter plot to examine the relationship between two numerical variables and the concentration of values between them. This improves the ability to recognize patterns and conveys a sense of depth. We improved our score from 79 to 81 percent by employing the Randomized Search CV as our hyper parameter adjustment. Figure 1 describes the Scatter Plot of the suggested system. To improve speed, data is dropped at the data preprocessing step after it has been collected and cleansed. To ensure that make predictions and accomplish the goal, we need develop the model in order to make the training data more universal before testing and training. Then, we should Use that understanding to fresh data that has never been seen before. To find out if the model is right, the next step is to train and test the method. Divide the data set into two sets in order to perform this: an exercise regimen as well as a testing set. We allocate 80% of the training set and 20% of the testing set to training. We fit the best model after testing and training to obtain a high accuracy score and Using machine learning methods, we forecast the price of the least expensive airline ticket.

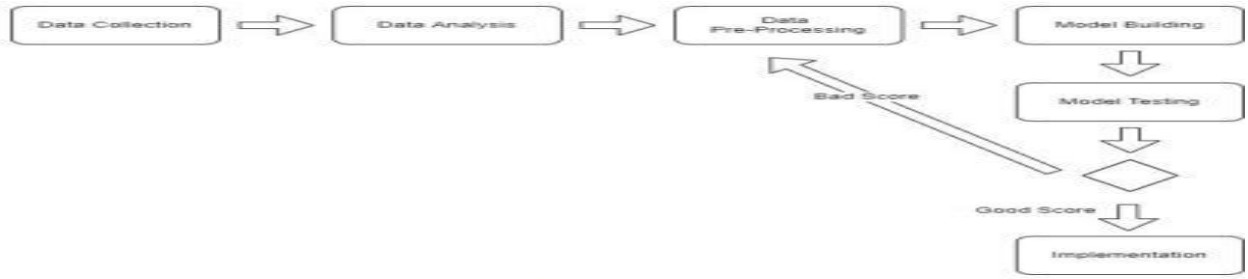

Fig. 1. Architecture of Proposed System

I. **ALGORITHM**

[1] The steps in the functioning technique are as follows:

[2] Step 1: From the training data set, select R observations at random.

[3] Step 2: Create classification trees for the selected sample points.

[4] Step 3: To display the system's categorization trees, choose a number A.

[5] Step 4: You'll repeat Steps 1 and 2.

[6] Step 5: Assign the samples to the class with the highest scores after determining the predictions for each decision tree's current data elements.

[7] The following are the stages involved in creating the

regression tree. $\sigma = \sqrt{\sum} (fv\ V\ v=1\ (\bar{Z}) - f\bar{)}$

II. **RESULTS**

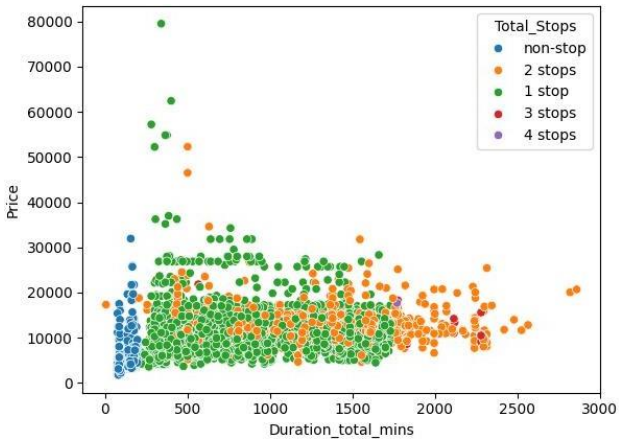

Fig.2.1. Scatter Plot 2 /V – 1

Random Forest Algorithm:

[8] 1. Pre-processing of data

[9] 2. Using the dataset for training with the Random Forest Algorithm

[10] 3. Projecting the study's outcome

[11] 4. Confusion Matrix Construction

[12] 5. Examining the potential outcome

[13] In order to tackle both regression [8] and classification tasks, Random Forest is an aggregate approach that combines several decision trees with an apparatus known as Bootstrap and Aggregation, commonly known as bagging.

[14] The random forest training approach trains tree learners using the conventional bootstrap aggregation technique, sometimes referred to as bagging. Using a training set Z = z1, zn and responses Y = y1,..., yn, After selecting a random

[9] When examining patterns and trends in airline fare prediction, scatter plots are an invaluable tool for visualizing interactions between two continuous variables in data analysis. Scatter plots make it easy to quickly identify any correlations or trends by graphing fare prices against variables like departure time, flight duration, or distance. They offer a user-friendly visualization of the distribution of the data, emphasizing any clusters or outliers that could have an effect on fare prediction algorithms. Furthermore, by displaying the kind and intensity of correlations between variables, scatter plots may help evaluate the applicability of other modelling approaches, such as linear regression. Scatter plots can provide important insights into the structure of the data, but they might not be able to adequately depict complicated interactions or nonlinear connections.

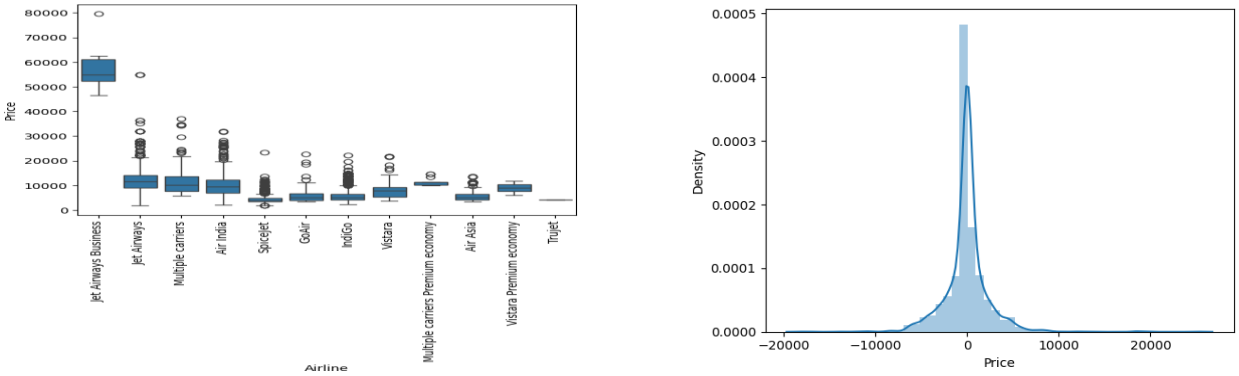

Fig.2.2 Box Plot

When examining airline fare data, boxplots may be a useful tool in exploratory data analysis since they can be used visualise the distribution, central tendency, and variability of numerical data. Key statistical metrics including the median, quartiles, and possible outliers are succinctly summarised by them. Boxplots can provide information on how fares are distributed among various airlines, routes, or time periods for the purpose of predicting flight fares. This

information can be used to spot trends and abnormalities. They might not, however, record intricate distributions within certain fare ranges or fluctuations over time. For a thorough understanding, more study and comparison with alternative visualisation approaches could be required. In conclusion, boxplots provide a simple means of visualising and contrasting flight fare distributions; however, they must be utilised in concert with other techniques for a comprehensive examination.

It demonstrates how to plot the test prediction using an imported randomised search CV and how to show the univariate data distribution in a hyper parameter Tweaking.

Fig 2.4 -Decision Regression Tree Regression trees and other decision trees are useful for predicting airline fares because of their interpretability, capacity to handle both numerical and categorical variables, and ability to capture intricate connections.

They can forecast prices by using variables like airline, route, and departure time efficiently, providing information on important aspects. However, methods like pruning and data quality enhancement should be used to overcome issues like overfitting and restricted extrapolation capabilities. Random Forests and other ensemble approaches can improve prediction accuracy, but for best results, thorough feature engineering and preprocessing are required. In general, decision trees provide a viable method for predicting travel costs by striking a balance between accuracy and comprehensibility.

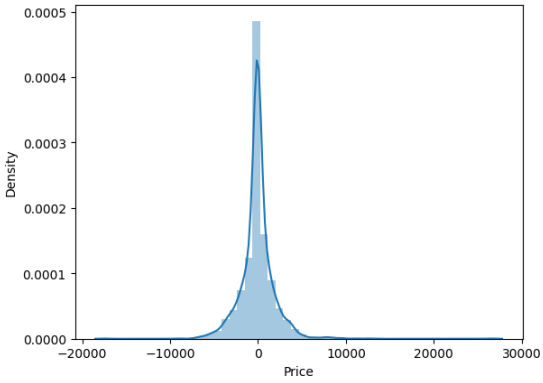

Fig.2.3 The Random Forest Distribution

III. CONCLUSION

In order to determine the lowest ticket price, we used machine learning models in our research work. Demand forecasting and ticket anticipation algorithms are used in literary reviews. We started out by providing an overview of airline pricing practices, which include recurrently changing ticket costs in reaction to internal and external factors variables. We explained the process by which airlines and customers work together to set dynamic ticket prices. To forecast the resultwe loaded the dataset and performed preliminary data analysis. Since the study of computer methods is known as machine learning, we transformed all text data kinds to numerical data types. We processed the category data and removed the useless columns. The datasets used for training and testing have been divided. This plan will be implemented to stop data leaks, and the Ensemble learning is utilized for both classification and regression. The Random Forest Regressor will be used to fit the model. After being imported, the data is preprocessed. Our hyperparameter optimization has allowed us to enhance the outcome. Our score increased from 79 to 82 percent by employing the Randomized Search CV as our hyper parameter adjustment.

REFERENCES

1. R. Ren, Y. Yang and S. Yuan, “Prediction of airline ticket price.” University of Stanford, 2014.
2. K. Tziridis, Kalampokas, A. G. Papakostas and I.
3. K. Diamantaras, “Airfare prices prediction using machine learning techniques.” In 2017 25th European Signal Processing Conference (EUSIPCO), August 2017, pp. 1036-1039, IEEE.
4. M. Papadakis, “Predicting Airfare Prices,” 2014

5. Gordiievych and I. Shubin, "Forecasting of airfare prices using time series," 2015 Information Technologies in Innovation Business Conference (ITIB), 2015, pp. 68-71, doi: 10.1109/ITIB.2015.7355055. [5] Y. Narangajavana, F.

6. J. Garrigos-Simon, J. S. García, S. ForgasColl, “Prices, prices and prices: A study in the airline sector.” Tourism Management, 41, 28-42, 2014

7. Lantseva, K. Mukhina, A. Nikishova, S. Ivanov, and

8. K. Knyazkov, “Data-driven modeling of airlines pricing.” Procedia Computer Science, 2015, 66, 267-276.

9. H. C. Wen and H. P. Chen, “Passenger booking timing for low-cost airlines: A continuous logit approach.” Journal of Air Transport Management, 64, 91-99, 2017.

10. S. J. Dominguez-Menchero, J. Riveraand E. Torres- Manzanera, “Optimal purchase timing in the airline market.” Journal of Air Transport Management, 40, 137-143, 2014.

11. M. Dedhia, A. Jadhav, R. Jagdale, and B. Palkar, “Optimizing Airline Ticket Purchase Timing.” Int. J. Recent Innovation Trends in Comput. Commun. (IJRITCC), 6(4), 296-298, 2018.