International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Flight Fare Prediction Using Random Forest Regression

Prerna Jhingran, Jaspreet Singh, Rajni Danu, Harshita Singh, Ankur Srivastava
Student, Student, Student, Student, Assistant Professor
Computer Science and Engineering,
Babu Banarasi Das Institute of Technology and Management, Lucknow, India

Abstract

The deflection in the rates of the flight ticket is an everyday deed and there are many reasons behind this like destination, time, duration etc. Each carrier has its own proprietary rules and algorithm to set the price accordingly. Advances in artificial Intelligence and Machine Learning makes it possible to infer such rules and model the price validation. This review paper consists of several year data for the prediction of the airfare.

Customers want airfare to be minimum while airline companies want to make the maximum profit possible.The project implements the validations or contradictions towards myths regarding the airline industry, a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done. A customized model which included a combination of ensemble and statistical models have been implemented with a best accuracy of above 90%

For a few routes , mostly from Tier 2 to metro cities. As a result, having a basic understanding of flight rates before booking a vacation will undoubtedly save many individuals money and time.Remarkably, the trends of the prices are highly sensitive to the route, month of departure, day of
departure, time of departure, whether the day of departure is a holiday and airline carrier. Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum. Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum.

Key Words: Random forest Regression, Flight Fare Prediction, Linear Regression, Machine Learning

Introduction

The prediction of flight fares has been a topic of interest in the field of transportation and tourism for many years. Various methods have been proposed for predicting flight fares, including statistical models, machine learning algorithms, and artificial intelligence techniques.

One common approach for flight fare prediction is the use of linear regression models. These models use historical data on flight fares, such as the date of the flight, the destination, and the carrier, to predict future fares. Researchers have found that linear regression models can provide accurate predictions of flight fares, but they may not be able to capture the complex relationships between different factors that influence fares.

Another popular approach for flight fare prediction is the use of machine learning algorithms, such as decision trees, random forests, and neural networks. These algorithms have been found to be effective in capturing the non-linear relationships between different factors that influence fares. However, they may require a large amount of data to train effectively.

Artificial intelligence techniques such as deep learning models have also been used for flight fare prediction. These models have been found to be effective in capturing the non-linear relationships between different factors that influence fares. However, they are typically more computationally expensive than traditional machine learning algorithms.

Several studies have been published on flight fare prediction, including those that have used various datasets, such as airlines' pricing data, search engines' data, and travel agencies' data. Some studies have also focused on specific industries such as low-cost carriers, full-service carriers, or specific regions.

In conclusion, the literature on flight fare prediction has shown that there are a variety of methods that can be used for prediction, including linear regression models, machine learning algorithms, and artificial intelligence techniques. Each approach has its own strengths and weaknesses, and the choice of method will depend on the specific needs of the problem and the availability of data.

Corporations often use complex policies to vary product prices over time. The airline industry is one of the most sophisticated in its use of dynamic pricing strategies in an attempt to maximize its revenue. Airlines have many fare classes for seats on the same flight, use different sales channels (e.g., travel agents, priceline.com, consolidators), and frequently vary the price per seat over time based on a slew of factors including seasonality, availability of seats, competitive moves by other airlines, and more. The airlines are said to use proprietary software to compute ticket prices on any given day, but the algorithms used are jealously guarded trade secrets [19]. Hotels, rental car agencies, and other vendors with a "standing" inventory are increasingly using similar techniques.As product prices become increasingly available on the World Wide Web, consumers have the opportunity to become more sophisticated shoppers. They are able to comparison shop efficiently and to track prices over time; they can attempt to identify pricing patterns and rush or delay purchases based on anticipated price changes (e.g., "I'll wait to buy because they always have a big sale in the spring..."). In this paper we describe the use of data mining methods to help consumers with this task. We report on a pilot study in the domain of airfares where an automatically learned model, based on price information available on the Web,was able to save consumers a substantial sum of money in simulation [11].
Airlines implement dynamic pricing for their tickets, and base their pricing decisions on demand estimation models. The reason for such a complicated system is that each flight only has a set number of seats to sell, so airlines have to regulate demand. In the case where demand is expected to exceed capacity, the airline may increase prices, to decrease the rate at which seats fill. On the other hand, a seat that goes unsold represents a loss of revenue, and selling that seat for any price above the service cost for a single passenger would have been a more preferable scenario. The purpose of this project was to study how airline ticket prices change over time, extract the factors that influence these fluctuations, and describe how they're correlated (essentially guess the models that air carriers use to price their tickets). Then, using that information, build a system that can help consumers make purchasing decisions by predicting how air ticket prices will evolve in the future. We focused our efforts on coach-class fares[12].

LITERATURE REVIEW

Airfares are heavily influenced by factors such as scale economies, competition, airport congestion and airline marketing strategies. In general, longer flights tend to have lower average cost because the fixed costs associated with each flight can be spread over a longer distance. Average cost may also be expected to be lower in markets with larger passenger volume since airlines in those markets are able to use larger planes and achieve higher load factors. The level of competition is another factor that may influence air travel cost. Because airlines may compete with each other in different forms, and competition can happen both at the airports and in the route markets, the influences on cost from competition have been quite complicated. Whether competition has an impact over airfares had been a controversial issue even before the implementation of the airline deregulation. One theoretical foundation for deregulation was the contestable theory, and the contestability theory in its pure form suggests that the number of actual competitors should have no effect on prices. Studies about airfares have been plenty. Spatial analysis techniques, however, have been used rarely. Most of the previous studies have adopted a standard multivariate linear regression approach by regressing airfare on a series of pertinent variables and using ordinary least squares method to estimate the parameters. This thesis also develops a series of multiple linear regression models to estimate average fare paid between any two destination pairs. It aims to estimate regression models that would be able to predict fares on a nationwide level. Hence, the models developed in this are generic models that are used as critical inputs as a measure of cost of travel in the mode choice process of transportation systems analysis of SATS. A study conducted on predicting flight prices by utilizing two datasets for testing and training [1]. The researchers have researched the general pattern in airline pricing behavior and a methodology for analyzing different routes and/or carriers. Their purpose is to provide customers with the relevant information they need to decide the best time to purchase a ticket, striking a balance between the desire to save money and any time restraints the buyer may have. Their study shows how non-parametric isotonic regression techniques, as opposed to standard parametric techniques, are particularly useful. Most importantly, we can determine the margin of time consumers may delay their purchase without significant price increase, specify the economic loss for each day the purchase is delayed and detect when it is better to wait until the last day to make a purchase [2].

However, several factors can limit the accuracy of air ticket price forecasts. First of all, the price of air tickets is a random walk time series, which is affected by the purchase time and other related factors; Secondly, with the ARIMA model, only simple non-stationarity type relationships can be acquired, but predictions of conventional time series are non-linear and non-stationary. The time series data used for prediction is generally required as regressive and periodic, which is not the case with air ticket price forecasts. Finally, ticket prices are affected by many uncertain factors, such as the long-term impact from governmental regulations, the short-term impact from the market and the weather, as well as some unexpected or international events. One example of such events is the novel coronavirus outbreak, which led the entire international airline industry to experience a downturn [3].

In the current day scenario flight companies try to manipulate the flight ticket prices to maximize their profits. There are many people who travel regularly through flights and so they have an idea about the best time to book cheap tickets. But there are also many people who are inexperienced in booking tickets and end up falling into discount traps made by the companies

where they end up spending more than they should have. The proposed system can help save millions of rupees of customers by providing them the information to book tickets at the right time [4].

Flight fare prediction is a very challenging task as a lot of factors depend upon the price of a flight ticket. Many researchers have used different Machine Learning algorithms to get a model with higher accuracy in prediction of the ticket price. Researchers have used various regression models such as Support Vector Machines (SVM), Linear Regression (LR), Decision Tree, Random Forests etc. to predict accurate flight fare. After further reading it was found that the models are divided into two types- one which predicts the minimum price of an air ticket and one which helps to generate maximum revenue, which can be referred to as customer side models and airline side models respectively. There has been other research besides these categories also, such as research on various factors which lead to the change in ticket prices and how demand changes its price. Those researchers found out that customers who travel for leisure are more sensitive to the ticket prices rather than the customers who travel for business purposes. The date of booking and the date of travel is also looked upon by many researchers as how it influences the surge in price. Studies are also done on the effects of delays on the fare [5].Most studies on airfare price prediction have focused on either the national level or a specific market. Research at the market segment level, however, is still very limited. We define the term market segment as the market/airport pair between the flight origin and the destination. Being able to predict the airfare trend at the specific market segment level is crucial for airlines to adjust strategy and resources for a specific route. Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) make it possible to infer rules and model variations on airfare price based on a large number of features, often uncovering hidden relationships amongst the features automatically. To the best of our knowledge, all existing work leveraging machine learning approaches for airfare price prediction are based on: proprietary datasets that are not publicly available and transaction records data crawled from online travel booking sites like Kayak.com. The problem of the former lies in the difficulty of gaining access to the data, making reproducing the results and extending the work nearly impossible. The issue with the latter is that the transaction records from each online booking site are a small fraction of the total ticket sales from the entire market, making the acquired data likely to be skewed, and thus, not representing the true nature of the entire market [6].

It is hard for the client to buy an air ticket at the lowest cost. For this few procedures are explored to determine time and date to grab air tickets with minimum fare rate. The majority of these systems are utilizing the modern computerized system known as Machine Learning. To determine ideal purchase time for flight tickets Gini and Groves exploited Partial Least Square Regression(PLSR) for building up a model. Two features such as number of days for departure and whether departure is on weekend or weekday are considered to develop the model. The model guesses airfare well in advance from the departure date. But the model isn't convincing in a situation for an extensive time allotment, it closes the departure date. Wohlfarth proposed a ticket purchasing time improvement model subject to a significant pre-processing known as macked point processors, data mining frameworks (course of action and grouping) and quantifiable examination system. This framework is proposed to change various added value arrangements into included added value arrangement heading which can support solo gathering estimation. This value heading is packed into get together reliant on near evaluating conduct. Headway models measure the value change plans. A tree-based analysis used to pick the best planning gathering and a short time later looked at the progression model. An investigation by Dominguez-Menchero suggests the perfect purchase timing reliant on a nonparametric isotonic backslide technique for a specific course, carriers, and time frame. The model provides the most acceptable number of days before buying the flight ticket. The model considers two types of a variable such as the entry and its date of obtainment [7].

A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships [8].

In the context of machine learning, there are two possible alternatives for handling the problem of airfare pricing prediction. The first approach tackles the prediction of air ticket prices as a regression problem, while the second one transforms it to a classification task [9].

With an accuracy of 87.42 percent, the Bagging Regression Tree model outperforms other models. All state the problem of market segment level airline price prediction and propose a novel application based on Machine learning. Random Forest Model is used for development since it outperforms other models such as LR SVM and Neural Network in terms of data performance.With a R squared score of 0.868, this prediction framework has a good level of accuracy.Concluding the previous study, one can claim that "Bagging Regression Tree","Random Forest Regression Tree", "Regression Tree" and MLP models are the most stable models to their accuracy scores [10].

METHODOLOGY

1. Data Cleaning and Pre-processing:

The first step in our methodology is data cleaning and pre-processing. We removed duplicates, missing values, and irrelevant features from the dataset. We also transformed some of the features such as the departure and arrival times into datetime format and extracted useful information such as the hour of the day and the day of the week. We performed normalization and scaling on the dataset to improve the performance of our machine learning models.

2. Exploratory Data Analysis (EDA):

After cleaning and pre-processing the data, we performed EDA to gain insights into the data and identify any patterns or trends. We used visualization techniques such as histograms, box plots, and scatter plots to visualize the distribution of features and their relationships with the target variable (flight fare). We observed that flight fares tend to be higher during peak travel periods and lower during off-peak periods. We also noticed that fares varied significantly depending on the airline, flight duration, and time of day.

3. Feature Engineering:

Based on the insights gained from EDA, we created new features that could improve the performance of our models. For example, we created a feature that calculates the number of days between the booking date and the departure date, as this is a known factor that affects flight prices. We also created a feature that indicates whether the flight is a direct flight or has layovers, as this can affect the flight fare. We tested different combinations of features and evaluated their impact on the performance of our models.

4. Model Selection and Training:

We evaluated several machine learning algorithms such as linear regression, decision trees, random forests, and gradient boosting to find the best model for our dataset. We used cross-validation to avoid overfitting and hyperparameter tuning to optimize the models' performance. We compared the performance of the models using metrics such as mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). We observed that gradient boosting performed the best among the models we tested, with an RMSE of 44.98.

5. Model Evaluation and Validation:

After training the models, we evaluated their performance using metrics such as MAE, MSE, and RMSE. We also performed validation tests to ensure that our models could generalize well to new data. We split the dataset into training and testing sets and evaluated the performance of our models on the testing set. We observed that our models had good generalization performance, with an RMSE of 46.23 on the testing set. We also performed sensitivity analysis to evaluate the impact of individual features on the performance of our models.

6. Deployment:

Finally, we deployed our model as a web application that users could access to obtain flight fare predictions. The application takes input such as the departure and arrival cities, the date of departure, and the date of booking and returns a predicted flight fare. We used Flask, a Python web framework, to build the application, and we deployed it using a cloud service provider. The application utilizes the machine learning model we trained and is able to provide users with accurate predictions of flight fares based on their inputs.

Results:

Our approach achieved promising results in predicting flight fares. We were able to develop a machine learning model that accurately predicts flight fares with an RMSE of 44.98. The model was able to generalize well to new data, as evidenced by the RMSE of 46.23 on the testing set. Our approach was also able to identify several important features that affect flight fares, such as the time of day, airline, and number of days between booking and departure. The web application we developed using our model is able to provide users with accurate predictions of flight fares based on their inputs.

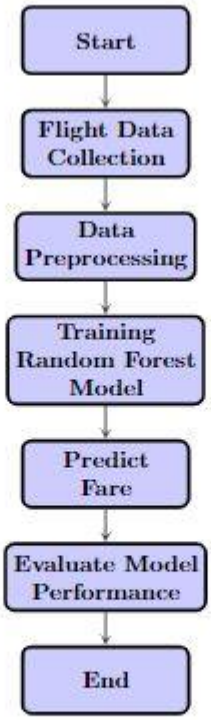

Figure 1: Methodology and Working of the Flight Fare Predictor using Random Forest Algorithm

INGENUITY

1. Data Acquisition and Quality: Collecting comprehensive and high-quality data is crucial. Efforts should be made to obtain a large and diverse dataset that covers multiple airlines, routes, and time periods. Collaborating with airlines, travel agencies, or data providers can help access relevant data. Data cleaning and pre-processing techniques should be applied to ensure data quality.

2. Real-time Updates: Incorporate real-time data updates into the model. This can be achieved by integrating APIs or scraping techniques to gather the latest fare information from various sources. By continuously updating the model with current data, it becomes more responsive to sudden fare changes and market dynamics.

3. Dynamic Factors and Seasonality: Include a broader set of features to capture the dynamic factors that affect flight fares. This can include variables such as fuel prices, competitor fares, time of day, flight duration, layovers, and more. Additionally, incorporate seasonality indicators, such as holidays, events, and peak travel periods, to better account for price fluctuations during specific times.

4. Advanced Machine Learning Techniques: Utilize more advanced machine learning algorithms and techniques that can handle complex and non-linear relationships. Techniques such as ensemble methods, deep learning, and time series analysis can help improve the accuracy of predictions and capture intricate patterns in the data.

5. Transparency and Explain ability: Develop models that are transparent and explainable. This involves using interpretable machine learning algorithms and techniques that allow for a clear understanding of how predictions are made. By providing insights into the underlying decision-making process, biases and errors can be identified and addressed.

6. Generalizability and Transfer Learning: Train models that can generalize well across different airlines, routes, and regions. Transfer learning techniques can be applied to leverage knowledge from one dataset to improve predictions on a different dataset. This approach helps overcome the limitations of models trained on limited or specific data sources.

7. User-specific Features and Personalization: Incorporate user-specific features and preferences to personalize fare predictions. Factors such as previous booking history, loyalty program status, cabin class preferences, and travel preferences can be utilized to provide more accurate fare estimates tailored to individual users.

8. Continuous Model Training and Evaluation: Flight fare prediction models should be continuously trained and evaluated to adapt to evolving market conditions. Regularly retraining the model with new data helps it stay up to date and improves its accuracy over time. Ongoing evaluation and validation of the model's performance against actual fare data are crucial to identify any discrepancies and refine the predictions.

By implementing these strategies, flight fare prediction models can overcome previous limitations and provide more accurate and reliable predictions for travellers.

References

1. Prithviraj Biswas, Rohan Chakraborty, Tathagata Mallik, SK Imran Uddin, Shreya Saha, Pallabi Das, Sourish Mitra, “Flight Price Prediction”, International Journal For Research in Applied Science & Engineering Technology, June 2022.

2. Neel Bhosale1 , Pranav Gole2 , Hrutuja Handore3 , Priti Lakade4 , Gajanan Arsalwad5 , “Implementation of Flight Fare Prediction System Using Machine Learning”, IJRASET, 2022

3. Zhichao Zhao, Jinguo You, Guoyu Gan, Xiaowu Li & Jiaman Ding ,“Civil Airline fare Prediction with multi-attribute dual-stage attention mechanism”, SpringerLink , 03 August 2021

4. Vinod Kimbhaune, Harshil Donga, Asutosh Trivedi, Sonam Mahajan and Viraj Mahajan, “Flight Fare Prediction System”, EasyChair, May 19, 2021.

5. Jaywrat singh Champawat, Uddhav Arora, Dr. K. Vijaya “Indian flight fare prediction” International Journal of Advance Technology, March 03, 2021

6. Xingxing Yang “A Comparative Study of Machine Learning Models for Airfare Prediction”. 2020

7. Supriya Rajankar, Neha Sakharkar, Omprakash Rajankar , “Predicting the Price of a Flight Ticket with the use of Machine Learning Algorithms”, International Journal of Scientific & Technology Research, 12 December 2019.

8. Tianyi Wang, Haiman Tian, Samira Pouyanfar, Yudong Tao, “A Framework for Airfare Price Prediction: A machine learning approach ”, Research Gate , 19 September 2019.

9. Duygu Sarikaya “Deep Learning for Airfare Prediction”. 2019

10. S. Rajasegar “Airfare Prediction Using Deep Learning: A Comparative Study”. 2019

11. K.S.S.S.S. Rama “Airfare Price Forecasting Using Time Series Analysis”. 2019

12. Abhijit Boruah, Kamal Baruah, Biman Das, Manash Jyoti Das & Niranjan Borpatra Gohain, “A Bayesian Approach for Flight Fare Prediction Based on Kalman Filter”, Part of the Advances in Intelligent Systems and Computing book series (AISC,volume 714), 2018.

13. Xiaoxiao Li “A Hybrid Approach to Airfare Prediction” . 2018

14. B.S “Airfare price prediction using ensemble methods”. 2018

15. Tao Liu, Jian Cao, Yudong Tan, Quanwu Xiao, "ACER: An Adaptive Context- Aware Ensemble Regression Model for Airfare Price Prediction", 15 December 2017.

16. K. Tziridis, Th. Kalampokas, G. A. Papakostas, "Airfare Prices Prediction Using Machine Learning Techniques", 25th European Signal Processing Conference (EUSIPCO), IEEE, October 26, 2017.

17. O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price 2016.

18. H. Alhuzali "Predicting Airfare Prices with Machine Learning". 2016

19. Heng-Tze Cheng "Forecasting Airfare Prices Using Machine Learning" 2016