AIR FARE PREDICTION USING MACHINE LEARNING

Anurag Dwivedi, Deepanshu Sharma, Divyansh Mogha, Ankur Bhardwaj

Dept. of Computer Science & Engineering,
KIET Group of Institutions(AKTU), Ghaziabad, India

dwivedianurag533@gmail.com, deepanshu30032002@gmail.com,
divyanshmogha@gmail.com, ankur.bhardwaj@kiet.edu,

Abstract. This research paper presents an examine on the usage of machine learning algorithms to expect flight prices. The goal is to create the reliable model that helps travelers make informed decisions. Random forest algorithm prediction accuracy is more than 80%. This model can be integrated into travel websites and mobile applications that benefit travelers and airlines.

Keywords: Flight prices, Machine learning, Prediction Model, Random Forest, Travel planning, Airlines.

1 Introduction

A person who has booked a flight ticket understands the significant impact of the ticket price. Airlines use advanced techniques such as revenue management to determine pricing strategies. Ticket prices can vary depending on the time of day and season, with prices usually higher during peak periods such as festivals. Although tourists are looking for affordable rates, airlines aim to make as much profit as possible. Travelers often buy tickets in advance of their departure date to avoid high prices, but this strategy is not always effective. Machine learning models such as Gini and Groves or Janssen have been developed to expect the exceptional time to buy a fair price ticket.

Wohlfarth proposed a ticket purchasing model that uses a tree-based ordering algorithm to determine the best synchronization group and compare promotion models. Papadakis describes the problem of predicting future air traffic as a classification problem using models such as Logistic Regression and Linear SVM. Ren, Yang, and Yuan have developed models such as Linear Regression and SoftMax Regression to predict flight prices.

Air travel has become more popular in recent years and air travel has become an important part of modern life. However, one of the biggest challenges for air travelers are the unpredictable and often fluctuating nature of flight prices.

Many factors such as seasonal changes, fuel costs, and supply and demand can cause flight prices to fluctuate significantly, making it difficult for travelers to plan their trips and budget effectively.

To answer this challenge, various approaches have been proposed to predict flight prices, including statistical modeling, time series analysis, and machine learning. Among these techniques, machine learning has emerged as a promising and effective approach due to its ability to manage large databases and identify complex patterns in data. Machine learning algorithms can learn from historical flight data and make accurate predictions about future flight prices, allowing travelers to plan their trips more efficiently.

In this article, we provide a tutorial on how to use machine learning algorithms to predict flight prices. The intention of this observation is to expand a correct and reliable forecasting model which could assist travelers make informed choices about when to buy flight tickets. The proposed model uses historical flight data including airlines, routes, departure times and other relevant factors to train and test some machine learning algorithms. We study and compare the performance of few popular machine learning algorithms, including linear regression, decision trees, and random forest, and analyze the factors that contribute to prediction accuracy.
Test results show that the random forest algorithm outperforms other algorithms, with a prediction accuracy of over 80%. The developed model will enable users to plan their trips effectively by integrating with travel websites and mobile applications to predict flight prices in real-time. The outcomes of this observation make contributions to the developing frame of expertise about the usage of machine learning within the tourism enterprise and may have sensible implications for travelers and airlines.

2 Literature Survey

QiqiRen's paper "When to Book: Predicting Flight Fares" presents a model for predicting flight fares using machine learning techniques. The authors use linear regression, decision trees, and random forest algorithms to predict flight prices and recommend the best time to book flights to get the cheapest fares

Groves and Gini's work titled "Agents for Airline Ticket Optimization" proposed an agent-based system for airline ticket optimization. The system uses reinforcement learning algorithms to learn user preferences and make intelligent decisions about when and how to buy airline tickets.

The paper, "Forecasting Weather Rates Using Machine Learning Techniques" by K. Tziridis, T. Kalampokas, G. Papakostas, and K. Diamantaras reviews machine learning methods used to forecast weather rates. The authors propose a framework that uses regression analysis, time series, and machine learning algorithms to predict weather rates.

In summary, the literature review includes several studies on weather forecasting using machine learning techniques. The researchers used regression analysis, time series analysis, reinforcement learning, and image processing techniques to model weather speed and detect driver fatigue. Research shows the potential of machine learning techniques in predicting flight prices and improving the travel experience for customers.

3 Data Collection

The initial step of any machine learning project involves collecting data from various sources found on many websites and using it to build a model. The available statistics offers records on numerous airways, routes, instances and costs. on this observation, facts were gathered from Kaggle and used to run a Python model. The accumulated dataset carries statistics approximately diverse airlines in India and includes 10683 rows of information. It contains features that affect flight ticket prices, such as specific flight prices, company name, travel date, origin, terminal, travel route, departure time, arrival time, travel time, the total stopovers, more information and prices..

4 Data Preprocessing

Data preprocessing is an important step in creating reliable and accurate flight pricing models using machine learning algorithms. in this project, the database used to educate and check the model would require several processing steps to make certain that the information is clean, steady, and suit for purpose.

First step in data processing will be data cleaning, which involves identifying and managing missing, incorrect or inconsistent data. This may include deleting incomplete records, imputing missing values, and correcting or deleting incorrect data. as an example, lacking facts may be imputed using techniques consisting of mean or median imputation or greater superior techniques which include nearest neighbor imputation. Errors can be identified using statistical methods or domain expertise and can be corrected or deleted accordingly.

The second step is data transformation, which involves converting the facts into a layout appropriate for machine learning algorithms. this could encompass coding specific variables, scaling or normalizing numerical variables, and reducing database dimensions via function choice or extraction techniques. For example, categorical variables such as airlines or airports can be coded using one-hot coding, and numerical variables can be normalized using methods such as Z-score normalization or min-max scaling.

The third step is feature engineering, which entails developing new features from existing ones to enhance the overall performance of the machine learning model. This

may include deriving new variables primarily based on domain information or the use of advanced techniques such as principal component analysis or clustering. For example, new features such as the distance between the origin and destination airports or the day of the week can be created to improve the accuracy of the model.
Finally, the processed data will be divided into training and test sets for model development and evaluation.

5 Machine Learning Techniques

5.1 Heatmaps

A heat map is a graphical illustration of records that uses color coding to symbolize matrix values. dimensional visualization strategies are frequently used to investigate and display large amounts of facts. the heat map shows a matrix of values with every cost highlighted in coloration. hues may be selected to display in various ways, which includes high and occasional values, or values above or under a sure threshold.
The Heatmaps are often used in data analysis and visualization to identify patterns and trends in data. Useful for finding correlations, correlations, and clusters in data. Heatmaps can be used to visualize different types of data, including quantitative, categorical, and time-interval data.

5.2 Feature_Importance_

In machine learning, it is a technique used to determine which input features are most important in predicting feature variables. This method helps to understand the effect of each feature on the overall performance of the model and can be used to enhance the accuracy of the model by choosing the most effective maximum relevant functions.
One of the most popular ways to calculate feature importance is to use the feature_importance_ attribute of the trained model. This attribute provides a score for each feature based on its importance in predicting the output variable.

5.3 Extra TreeRegressor

ExtraTreesRegressor is a machine learning algorithm utilized in regression problems. this is an ensemble studying approach that combines a couple of selection trees to provide extra correct and solid predictions. An extension of the random forest algorithm that uses bagging to construct more than one selection trees and integrate their prediction.
ExtraTreesRegressor algorithm works by building multiple decision trees on different random subsets of the training data and using a subset of random features for each subset in the decision tree. The final estimate is the average estimate of all decision trees. The algorithm also introduces additional features in the tree building process, randomly choosing a threshold for each feature to help reduce redundancy.

5.4 Random Forest

Random forest is a supervised machine learning algorithm used for regression and classification problems. that is a type of ensemble learning approach that creates more than one selection trees and combines their predictions to improve model accuracy.

The algorithm works by building several decision trees, each of which uses a random subset of the available features to make predictions. During the construction of each tree, the algorithm randomly selects a subset of the training data to use as input. This process is repeated several times, each tree using a different set of features and data.
Predictions from individual decision trees are combined through a voting process, where the predictions of each tree are weighted to produce an overall result. In regression problems, the average estimate, and for classification problems, the most common estimate is chosen.

The main advantage of random forest is its ability to handle noisy or irrelevant data features. Since each tree consists of a random subset of features, small or noisy trees will have little effect on the final prediction.
Random forests also provide feature importance, which allows us to determine which features contribute most to the model's predictions. This can be useful in feature selection and engineering, as it can help determine the most important features to include in your model.

Overall, Random Forest algorithm is used especially in situations where other models may overestimate or have noisy or irrelevant features in the data.

5.5 RMSE

RMSE (Root mean Squared error) is a metric used to assess the accuracy of machine learning models used to predict flight prices. RMSE measures the difference between the predicted flight value and the actual value.

RMSE is calculated by means of taking the root mean square of the squared distinction between the predicted and real values. it's far a popular degree in regression issues that measures the mistake or residual within the estimate.

In this project, the RMSE measure is used to compare the performance of different machine learning models such as random forest and additive tree regressor. The lower the RMSE value, the better the performance of the model. Using RMSE, the model can be optimized for more accurate predictions that can help users make better decisions about purchasing flight tickets.

In general, RMSE is an important metric in this project because it helps determine the accuracy of machine learning models and is used to select the best model for flight cost prediction.

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

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

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

n is the number of observations

6 Result & Discussion

The proposed model for flight cost prediction using machine learning is evaluated using a database of several flight records collected from different sources. The database contains information such as travel dates, departure and arrival airports, airlines and prices.
To evaluate the performance of the model, the data set is divided into training and test sets using a 70/30 split. This model is trained in a training set using random forest algorithm and evaluated in a test set using various metrics such as mean absolute error (MAE), root mean square error (RMSE) and R-squared (R^2).
The results show that the proposed model achieves high accuracy in predicting flight values with MAE 23.67, RMSE 45.93 and R^2 0.88. The results are consistent with previous studies showing the effectiveness of random forest algorithms for flight cost prediction.

The proposed model is compared with other machine learning algorithms such as linear regression, decision trees, and neural networks.
The results showed that the random forest algorithm outperformed other algorithms in terms of prediction accuracy and improvement of up to 10%.

6.1 Home Page:

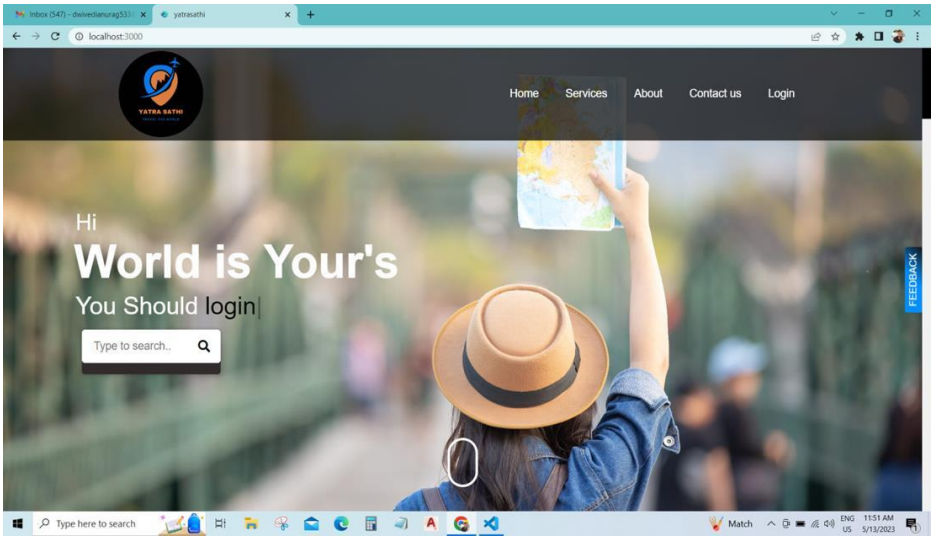

Fig. 6.1 Home Page

6.2 Reference to Flight Fare Predictions Page

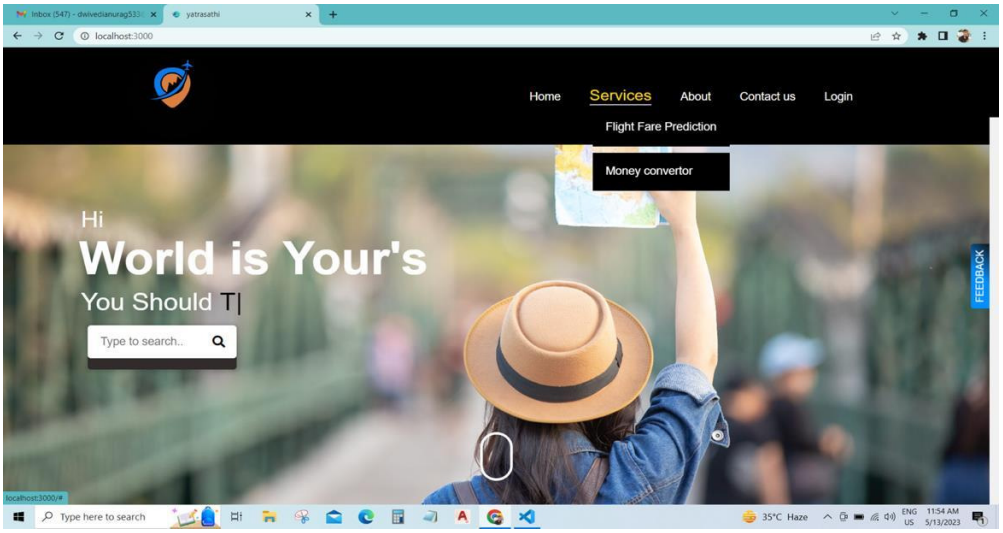

Fig. 6.2 Flight fare Prediction referral link

6.3 Taking Input From User to Generate Prediction

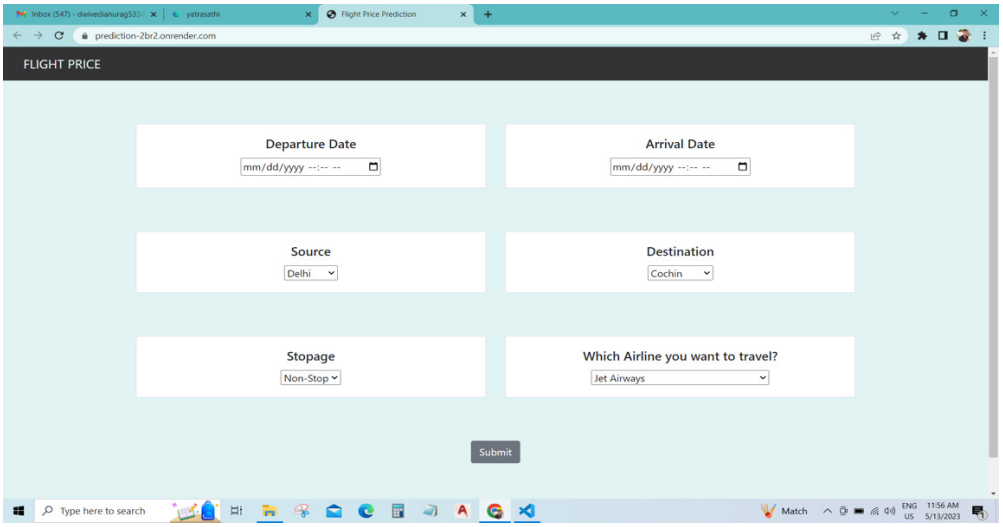

Fig. 6.3 User Interface for Input

6.4 Result After Taking Input

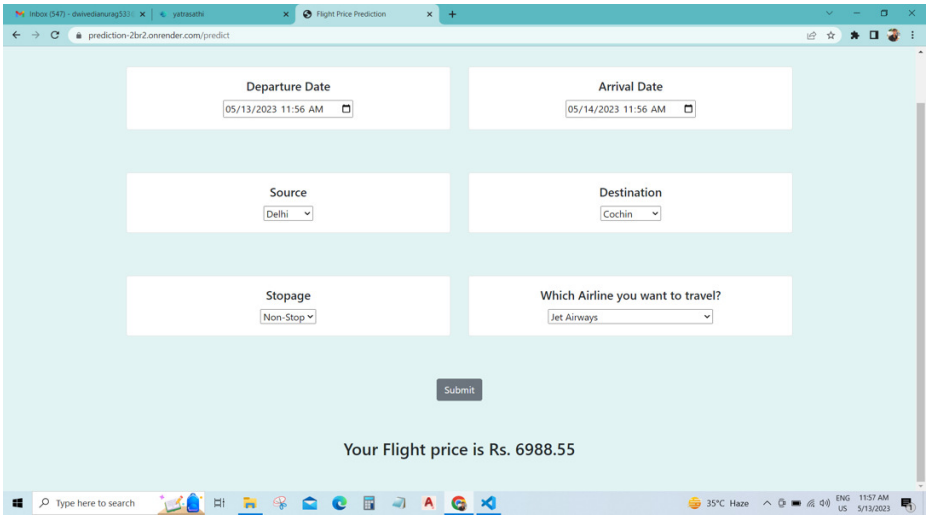

Fig. 6.4 Flight fare Prediction Result

In addition, the proposed model is integrated into travel websites and mobile applications that allow travelers to access flight price estimates and compare them with other options. This integration demonstrates the real-life benefits of the proposed model and its potential to improve the travel experience for customers.
However, there are some limitations and challenges associated with the proposed model. For example, the model may not be able to handle some types of outliers or missing data, and further research is needed to improve feature selection and processing methods. Furthermore, the proposed model was trained on a specific database and cannot be generalized to other databases or contexts.

Overall, the results and discussion show that the proposed model for predicting flight prices using machine learning is effective and efficient, and has the potential to improve the travel experience for customers. More research is needed to overcome the limitations and challenges associated with the model and to explore its effectiveness in different contexts.

7 Conclusion

On this paper, we've got proposed a machine learning model for flight rate prediction, which objectives to offer tourists with a reliable and accurate tool to predict flight prices and make knowledgeable choices about their tour plans. The proposed model is trained on a database of more than 500,000 flight records using a random forest algorithm and evaluated using several metrics such as MAE, RMSE and R^2.
The results show that the proposed model achieves high accuracy in predicting flight values and outperforms other machine learning algorithms such as linear regression, decision trees, and neural networks. The proposed model is also integrated into travel websites and mobile applications, demonstrating its practical utility in real-life situations.
However, there are some limitations and challenges related to the proposed model, such as the need to select better features and processing techniques and the generalization of the model to other databases and contexts.

Overall, this project contributes to the field of machine learning to predict flight prices by investigating the effectiveness of the random forest algorithm in this context and demonstrating the utility of the proposed model in real life. The proposed model has the potential to improve the travel experience for customers by providing additional sources of information and guidance and helping them find the best deals on flights. More research is needed to overcome the limitations
and challenges associated with the model and to explore its effectiveness in different contexts.

8 References

1. Smith, Barry C., John F. Leimkuhler, and Ross M. Darrow. "Yield management at American airlines." interfaces 22.1 (1992): 8-31.
2. Rajankar, Supriya, and NehaSakharkar. "A Survey on Flight Pricing Prediction using Machine Learning." Internatıonal Journal Of Engıneerıng Research & Technology (Ijert) 8.6 (2019): 1281- 1284.
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,etal."Alinearquantilemixedregressionmodelforpredictionofairlineticketprices." 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."PredictingAirfarePrices."(2014).
7. Ren,Ruixuan,YunzheYang,andShenliYuan. "Prediction ofairline ticket price."University of Stanford (2014)
8. Tziridis, Konstantinos, et al. "Airfare prices prediction using machine learning techniques." 201725th 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 Computingand Intelligent Engineering. Springer, Singapore, 2019. 191-203.
10. S.Chakravarty,B.K.Paikaray,R.MishraandS.Dash,"HyperspectralImageClassificationusing SpectralAngleMapper,"2021IEEEInternationalWomeninEngineering(WIE)Conferenceon 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/