ML-POWERED WEB APPLICATION FOR AIRFARE PREDICTION

B. Rahmani [1], D. B. Pallerla [2], H. S. Munjampally [2], N. C. Bavirisetti [2], N. Reddy Yalamacharla [2], R. Donadula [2], A. Maazallahi [3], M.A. Salari [2], P. Norouzzadeh [4], D. Niknami [5]

[1] School of Medicine-HCOR, Saint Louis University, Saint Louis, MO
[2] Department of Computer Science, Saint Louis University, Saint Louis, MO
[3] University of Tehran, Tehran, Iran
[4] Department of Professional Studies, Saint Louis University, Saint Louis, MO
[5] Islamic Azad University, Qazvin, Iran

ABSTRACT

This research investigates the prediction of flight ticket prices using advanced machine learning techniques. Utilizing a dataset of 10,684 records with 11 features including spanning departure and arrival times, number of stops, ticket prices, and airline details. The study implements various predictive models, including Linear Regression, Decision Trees, K-Nearest Neighbors (KNN), and Random Forest. Comprehensive feature engineering and selection methods identified key variables such as total stops, flight duration, and airline type as significant predictors of ticket prices. Among the evaluated models, the Random Forest algorithm demonstrated superior performance, achieving a Root Mean Squared Error (RMSE) of 3843.99 and an R^2 score of 0.80, significantly outperforming other approaches. Additionally, the model was deployed in a Flask web application for real-time predictions, providing a practical tool for travelers and industry stakeholders. The findings emphasize the efficacy of ensemble methods in capturing complex relationships within flight pricing data, offering valuable insights for optimizing budget planning and improving user satisfaction.

1. INTRODUCTION

The cost of air travel is influenced by a complex interplay of factors, including the number of stops, airline policies, departure and arrival times, and the source and destination of the journey. Understanding these dynamics is critical for optimizing ticket pricing strategies and enhancing customer satisfaction. Accurate flight price prediction can empower travelers with better budget planning and provide service providers with data-driven insights for competitive pricing. This study focuses on leveraging machine learning techniques to predict flight ticket prices using a rich dataset of flight bookings between major Indian metro cities. By analyzing 10,684 records and 11 features, including both numerical and categorical variables, the research identifies key factors that significantly impact flight pricing. Advanced feature selection methods, such as correlation analysis and the ExtraTreesRegressor algorithm, are employed to ensure that the most relevant predictors are incorporated into the models. Through the implementation of multiple machine learning models including Linear Regression, Decision Trees, Random Forest, K-Nearest Neighbors, and Naïve Bayes; this study evaluates the predictive performance of each approach. The Random Forest model emerged as the most effective, capturing complex relationships within the data and achieving the highest accuracy. This paper highlights the practical utility of these models in addressing the challenges of flight price prediction, with potential applications in real-time pricing tools and customer decision-making support systems.

DOI:10.5121/caij.2025.12301

Dr. Nadeem and his colleagues compare multiple algorithms, including linear regression, decision trees, random forests, and XGBoost, concluding that random forests outperform others in predicting flight ticket prices [1] . Another study analyzes flight fare prediction using essential exploratory data analysis techniques and various features such as airline type, arrival time, departure time, and flight duration [2]

"Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting" research introduces a Gated Recurrent Unit (GRU) model, demonstrating its effectiveness in capturing temporal dependencies in flight fare data [3] . Dr. wang proposes a novel application of machine learning for airfare price prediction, focusing on modeling price variations based on many features [4].

"An Investigation into Predicting Flight Fares in India Using Machine Learning" study explores several machine learning models to predict airfares based on multiple characteristics, enhancing flight price prediction accuracy [5]. Other research investigates the use of machine learning models like Linear Regression, Decision Tree Regressor, and Random Forest Regressor to predict flight fares, emphasizing feature engineering and model evaluation [6].

"A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques" paper analyzes airfare price prediction by extracting effective features from a large dataset and applying AI techniques to find similarities in pricing policies of different airlines [7] . Kalampokas and his group discussed predictive performance on flight pricing data using Spark ML models, focusing on Random Forest, Gradient Boosted Trees, Decision Trees, and Factorization Machines [8]

"Model Monitoring and Dynamic Model Selection in Travel Time-series Forecasting" research introduced a data-driven framework to continuously monitor and maintain deployed time-series forecasting models' performance, validated on flight and hotel price datasets [9] . Other article combines data from multiple sources, including competitor data and pricing, to study customer behavior and improve booking predictions using machine learning [10]. Dr. Lu compared multiple algorithms, including linear regression, decision trees, random forests, and XGBoost, concluding that random forests outperform others in predicting flight ticket prices [11].

Another paper discusses the issue of airfare prediction by selecting a set of characteristics that define a typical flight, assuming these characteristics influence the price of an airline ticket. It explores how flight ticket prices fluctuate based on parameters such as flight schedule, destination, duration, and various occasions like vacations or the holiday season. The study applies machine learning algorithms to predict these fluctuations [12].

"Aircraft Ticket Price Prediction Using Machine Learning" research focuses on predicting flight ticket prices by collecting data on specific air routes, including features like departure time, arrival time, and airlines over a specific period. It applies machine learning regression methods to predict prices at given times, aiming to assist frequent travelers in obtaining the best deals [13].

Previous studies often rely on limited datasets, either geographically or temporally constrained, which may not capture diverse pricing patterns. This study highlights the significance of key variables, such as total stops and flight duration, and provides visualizations like decision trees and heatmaps to enhance interpretability.

2. DATA DESCRIPTION

Dataset used for this project contains flight booking data from major Indian metro cities, sourced from Kaggle1. It includes detailed information about flights such as departure times, arrival times, ticket prices, and the number of stops. These features are crucial in understanding the

dynamics of flight pricing. The primary objective was to use this data to identify patterns and build machine learning models that predict ticket prices accurately.

Figure 1 shows the correlation between features of data. This heatmap visualizes the correlation between dataset features and the target variable, ticket price. Strong correlations are observed for features such as the number of stops (0.603), flight duration (0.509), and specific airline types (e.g., Jet Airways, 0.416), indicating their significant influence on price prediction.

3. METHODOLOGY

To identify the most effective machine learning approach for flight price prediction, several models were implemented and evaluated:

3.1. Linear Regression

Linear Regression served as the baseline model. It assumes a linear relationship between the features and the target variable (ticket price). While it is interpretable and computationally efficient, its failure to capture non-linear relationships in the dataset limits its performance.

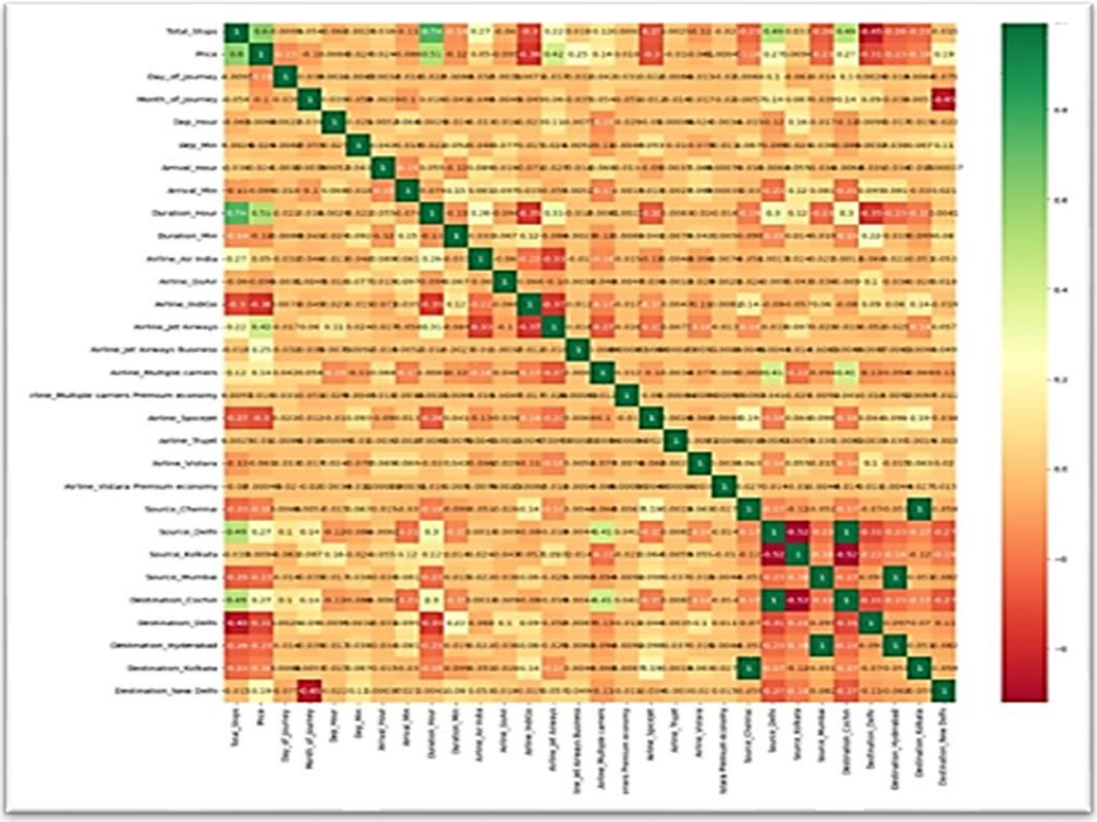

Figure1: Correlation Heatmap

3.2. Decision Trees

Decision Trees create a flowchart-like structure where features are split at nodes to predict the target variable. They are efficient at capturing non-linear relationships and provide interpretability by visually representing the decision process. However, they are prone to overfitting if not properly tuned.

3.3. Random Forest

Random Forest, an ensemble learning method, combines multiple decision trees to improve prediction accuracy and reduce overfitting. By aggregating predictions from each tree, Random Forest achieves robust performance and handles complex feature interactions effectively. This model emerged as the best-performing approach in this study.

3.4. K-Nearest Neighbors (KNN)

KNN predicts the ticket price by considering the prices of the nearest data points in feature space. It is simple to implement and works well for smaller datasets but becomes computationally expensive as the dataset size increases. Its performance depends heavily on the choice of the hyperparameter k.

3.5. Naïve Bayes

Naïve Bayes is a probabilistic model based on Bayes' theorem, assuming independence between features. While it is fast and efficient for classification tasks, its assumptions of feature independence can limit its accuracy for datasets with correlated variables like flight pricing.

4. RESULTS

Based on Table 1, Random Forest, with almost %80 accuracy, has the best performance. Decision Tree as a part of Random Forest has second place.

Table 1: Model Comparison

Model	RMSE	R^2 Score	Accuracy
Linear Regression	4600	0.67	Not Applicable
Decision Tree	4200	0.74	75.81%
Random Forest	3843.99	**0.80**	**78.42%**
K-Nearest Neighbors	**4800**	0.60	66.23%
Naïve Bayes	Not Applicable	Not Applicable	72.45%

Figure 2 depicts the importance of various features in predicting flight ticket prices. The most influential features include the total number of stops, day of the journey, flight duration, and airline type. These features exhibit the highest contribution to the variability in ticket pricing, guiding the model toward more accurate predictions. These feature selection methods ensured that only the most relevant and influential variables were included in the modeling process, improving the efficiency and accuracy of the predictive models.

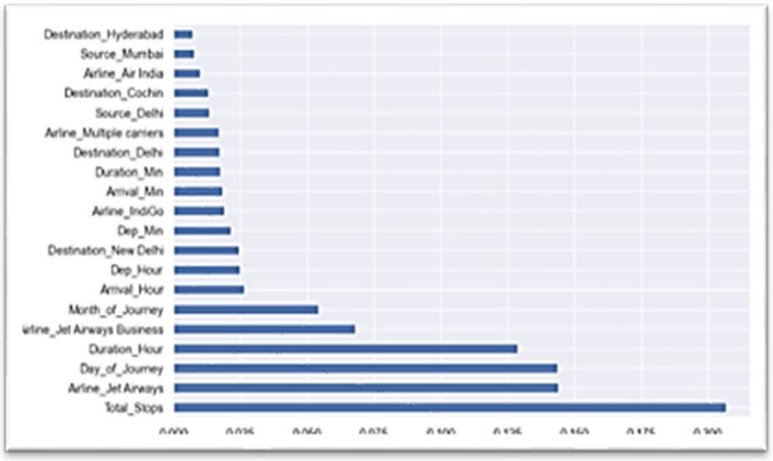

Figure 2: The importance of features

5. CONCLUSION

The results of this study demonstrate the efficacy of machine learning models in predicting flight ticket prices, with Random Forest emerging as the most reliable approach. By leveraging its ensemble learning capabilities, Random Forest successfully captured complex relationships between features such as total stops, duration hour, and airline type. Feature selection played a pivotal role in improving model performance. Techniques like correlation heatmaps and ExtraTreesRegressor effectively identified the most impactful features, ensuring that the models focused on the variables that truly drive ticket prices.

6. WEB APPLICATION

Figure 3 presents the final model used to develop a sample web application. The final model is integrated into a Flask web application. Users can input flight details such as date, departure, arrival times, stops, airline, source, and destination. The application predicts the flight price based on the trained RandomForestRegressor model. Users receive the predicted flight price as the output.

Model Name: RandomForestRegressor
X Features: Total_Stops, Day_of_Journey, Month_of_Journey, Dep_Hour, Dep_Min, Arrival_Hour, Arrival_Min, Duration_Hour, Duration_Min, Airline (Multiple binary-encoded columns), Source (Multiple binary-encoded columns), Destination (Multiple binary-encoded columns)
Y Output: Predicted Flight Price
Performance Metrics:
Mean Absolute Error (MAE): 1312.32
Mean Squared Error (MSE):4685488.41
Root Mean Squared Error (RMSE): 2064
R-squared (R2) Score: 0.80

Figure 3: Final Model

Figure 9 shows a sample web application which offers a user-friendly interface for obtaining quick predictions, enhancing user experience in predicting flight prices2. This web application works for Delhi, Kolkata, Mumbai, and Hyderabad. Figure 4 is an example of a flight from Delhi to Kolkata. The actual price is ₹10401 ($123), and the predicted price is ₹8748 ($103) with %84 accuracy.

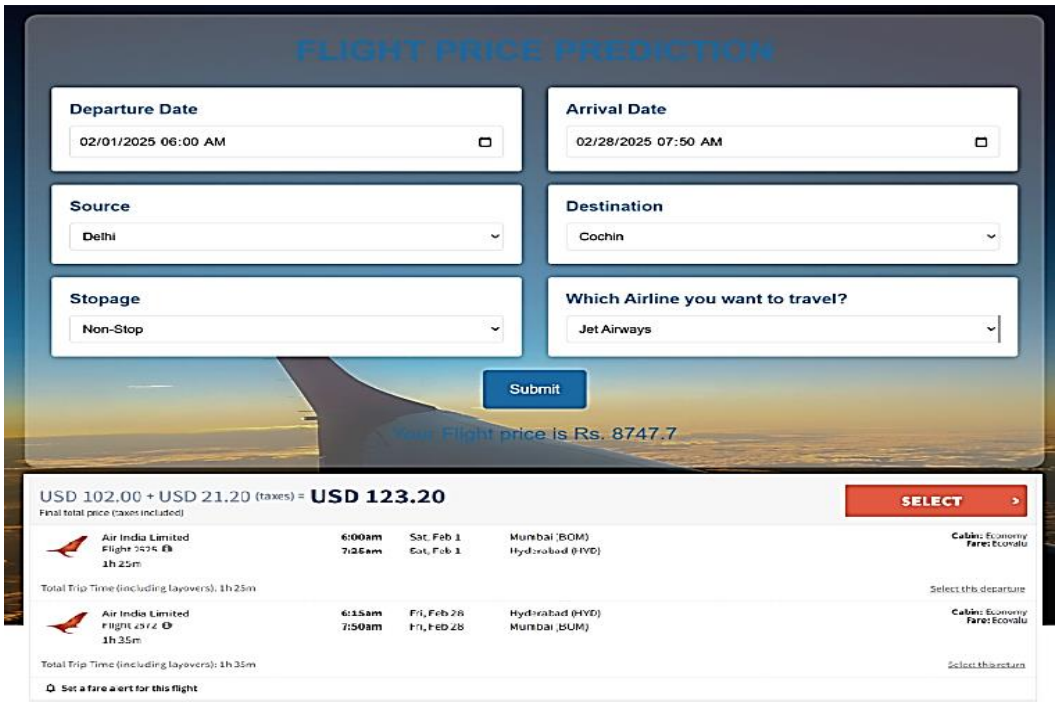

Figure 4: Flight Fare Prediction-Web Application

REFERENCES

[1] R. Nadeem and T. Sivakumar, "Flight Fare Forecasting: A Machine Learning Approach to Predict Ticket Prices," in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-981-99-3878-0_60.

[2] A. Panigrahi, R. Sharma, S. Chakravarty, B. K. Paikaray, and H. Bhoyar, "Flight Price Prediction Using Machine Learning," in CEUR Workshop Proceedings, 2021. doi: 10.55041/ijsrem25823.

[3] W. A. Degife and B. S. Lin, "Deep-Learning-Powered GRU Model for Flight Ticket Fare Forecasting," Applied Sciences (Switzerland), vol. 13, no. 10, 2023, doi: 10.3390/app13106032.

[4] T. Wang et al., "A framework for airfare price prediction: A machine learning approach," in Proceedings - 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI 2019, 2019. doi: 10.1109/IRI.2019.00041.

[5] V. Lal, P. Stynes, and C. H. Muntean, "An Investigation into Predicting Flight Fares in India Using Machine Learning Models," in Lecture Notes in Networks and Systems, 2023. doi: 10.1007/978-3-031-42317-8_9.

[6] K.D.V.N.Vaishnavi, L. Hima Bindu, M. V Satwika, K. Udaya Lakshmi, M. Harini, and N. Ashok, "FLIGHT FARE PREDICTION USING MACHINE LEARNING," EPRA International Journal of Research & Development (IJRD), 2023, doi: 10.36713/epra14763.

[7] T. Kalampokas, K. Tziridis, N. Kalampokas, A. Nikolaou, E. Vrochidou, and G. A. Papakostas, "A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3274669.

[8] P. Wong, P. Thant, P. Yadav, R. Antaliya, and J. Woo, "Using Spark Machine Learning Models to Perform Predictive Analysis on Flight Ticket Pricing Data," arXiv preprint arXiv:2310.07787, 2023.

[9] R. Candela, P. Michiardi, M. Filippone, and M. A. Zuluaga, "Model Monitoring and Dynamic Model Selection in Travel Time-Series Forecasting," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2021. doi: 10.1007/978-3-030-67667-4_31.

[10] D. Hopman, G. Koole, and R. van der Mei, "A machine learning approach to itinerary-level booking prediction in competitive airline markets," International Journal of Revenue Management, vol. 12, no. 3–4, 2021, doi: 10.1504/IJRM.2021.120347.

[11] J. Lu, "Machine learning modeling for time series problem: Predicting flight ticket prices," arXiv preprint arXiv:1705.07205, 2017.

[12] R. R. Subramanian, M. S. Murali, B. Deepak, P. Deepak, H. N. Reddy, and R. R. Sudharsan, "Airline Fare Prediction Using Machine Learning Algorithms," in Proceedings - 4th International Conference on Smart Systems and Inventive Technology, ICSSIT 2022, 2022. doi: 10.1109/ICSSIT53264.2022.9716563.

[13] J. Mukane, S. Pawar, S. Pawar, and G. Muley, "Aircraft Ticket Price prediction using Machine Learning," Int J Res Appl Sci Eng Technol, vol. 10, no. 2, 2022, doi: 10.22214/ijraset.2022.40253.