IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 7, March 2026

Impact Factor: 8.2

A Machine Learning Framework for Dynamic Airline Fare Forecasting

P. Rithin Rajpoot[1], D Mohith[2], K Kavya[3], Dr Punyaban Patel[4]

UG Scholars, Department of Computer Science & Engineering[1,2,3]

Professor, Department of Computer Science & Engineering[4]

CMR Technical Campus, Hyderabad, India

Abstract: *Airline corporations employ complex dynamic pricing strategies influenced by financial, marketing, commercial, and social factors, making it difficult for consumers to purchase tickets at the lowest possible price. This paper addresses the problem of airfare price prediction using machine learning (ML) techniques. A set of flight features is selected and eight state-of-the-art ML models are applied and compared for prediction accuracy. A novel dataset of 1,814 flight records from Aegean Airlines (Thessaloniki to Stuttgart) is used for training and testing. The experimental results demonstrate that ML models can handle this regression problem effectively. Among all models tested, Regression Tree achieved the highest accuracy of 0.99, closely followed by Random Forest and Bagging Regression at 0.92. The study also investigates feature importance and the dependency of accuracy on the selected feature set.*

Keywords*:* Airfare Prediction, Machine Learning, Regression Tree, Random Forest, Multilayer Perceptron (MLP), Extreme Learning Machine (ELM), Support Vector Machine (SVM), Bagging Regression, Feature Selection, Dynamic Pricing

I. INTRODUCTION

Now-a-days, airline corporations use complex strategies and methods to assign airfare prices dynamically. These strategies take into account several financial, marketing, commercial, and social factors closely connected with final airfare prices. Due to the high complexity of pricing models applied by airlines, it is very difficult for a customer to purchase an air ticket at the lowest price, since the price changes dynamically. For this reason, several techniques have been proposed to provide the right time for the buyer to purchase an air ticket by predicting the airfare price [1], [2].

The majority of these methods make use of sophisticated prediction models from the computational intelligence research field known as Machine Learning (ML). Groves and Gini applied a PLS regression model to optimize airline ticket purchasing with 75.3% accuracy. Papadakis predicted if the price of the ticket would drop in the future using Ripple Down Rule Learner (74.5%), Logistic Regression (69.9%), and Linear SVM (69.4%) models. Janssen proposed a linear quantile mixed regression model to predict air ticket prices with acceptable performance for cheap tickets many days before departure. Ren, Yang, and Yuan studied Linear Regression (77.06%), Naive Bayes (73.06%), Softmax Regression (76.84%), and SVM (80.6%) models [3], [4].

The contribution of this paper is summarized as follows: (1) airfare price prediction using a comprehensive ML study, (2) investigation of feature influence on airfare prices, and (3) performance analysis of eight state-of-the-art ML models. The paper applies and compares MLP, ELM, Random Forest, Regression Tree, Bagging Regression, Polynomial SVM, Linear SVM, and Linear Regression. A novel dataset of 1,814 flight records from Aegean Airlines is constructed for experiments, and comprehensive feature analysis is conducted to determine which flight characteristics most significantly affect prediction accuracy [5], [6].

DOI: 10.48175/IJARSCT-32213

II. LITERATURE SURVEY

Research on airline pricing and ticket purchasing strategies has explored a range of analytical and machine learning approaches to understand fare dynamics and optimize buying decisions. Malighetti et al. [1] analyzed the pricing policies of Ryanair using a year's fare data and hyperbolic price functions to estimate optimal pricing curves for each route. Their

analysis revealed a positive correlation between average fare, route length, flight frequency, and the percentage of fully booked flights.

In the context of predictive modeling, Groves and Gini [3] proposed a regression-based approach to estimate expected future prices and assess the risk of price fluctuations, demonstrating that guided purchase strategies can reduce costs within the two months prior to departure. An intelligent agent was later introduced [4] to autonomously optimize ticket purchase timing, achieving performance closer to the optimal purchase policy than traditional decision-theoretic methods.

Janssen [6] developed multiple regression models, including Linear Quantile Mixed Regression, using a large dataset of flight prices, showing that these models effectively capture price behavior well in advance of departure. Machine learning techniques such as Linear Regression, Naive Bayes, Softmax Regression, and Support Vector Machines were applied by Ren et al. [7], with SVM achieving the highest accuracy of 80.6%, highlighting the effectiveness of machine learning in airfare prediction.

III. PROPOSED METHODOLOGY

A. Proposed System

This paper deals with the problem of airfare price prediction using a comprehensive set of ML models. A set of features characterizing a typical flight is determined, supposing that these features directly affect the price of an air ticket. The features are applied to eight state-of-the-art ML models, the performance of each model is compared, and a comprehensive feature analysis is conducted to determine which flight characteristics most significantly affect prediction accuracy.

B. System Architecture

The system architecture follows a structured pipeline from raw data collection through preprocessing, feature extraction, model training, and evaluation. The pipeline begins with dataset upload and ingestion, followed by preprocessing steps that handle missing values and encode categorical variables. The processed data is then fed into eight parallel ML model training pipelines. Finally, outputs are evaluated and accuracy results are visualized for comparison.

C. System Modules

The system consists of four primary modules. (1) Upload Airfare Prices Dataset: The dataset is uploaded to the application for processing. (2) Pre-process Dataset: Missing values are removed and non-numeric string values are encoded to numeric data using Label Encoder. (3) Run ML Algorithms: The dataset is split 80/20 into training and testing sets; each of the eight algorithms is trained and evaluated for prediction accuracy. (4) Accuracy Comparison Graph: An accuracy comparison graph is plotted across all algorithms, with algorithm names on the X-axis and accuracy values on the Y-axis [7], [8].

D. Machine Learning Algorithms Used

Multilayer Perceptron (MLP) is a feedforward artificial neural network that maps input features to outputs through multiple hidden layers using backpropagation [8]. Extreme Learning Machine (ELM) uses randomly assigned hidden layer weights and learns only output weights, enabling fast training with competitive accuracy [10]. Random Forest constructs multiple decision trees and outputs the mean prediction, reducing overfitting [14]. Regression Tree partitions the feature space into regions and predicts a constant value for each, yielding high accuracy and interpretability [15].

DOI: 10.48175/IJARSCT-32213

Bagging Regression trains multiple models on random subsets to improve stability and reduce variance. Support Vector Machine with Polynomial and Linear kernels finds the optimal hyperplane for price categorization [11]. Linear Regression models the direct linear relationship between input features and airfare price and serves as the baseline comparator for all other models in this study.

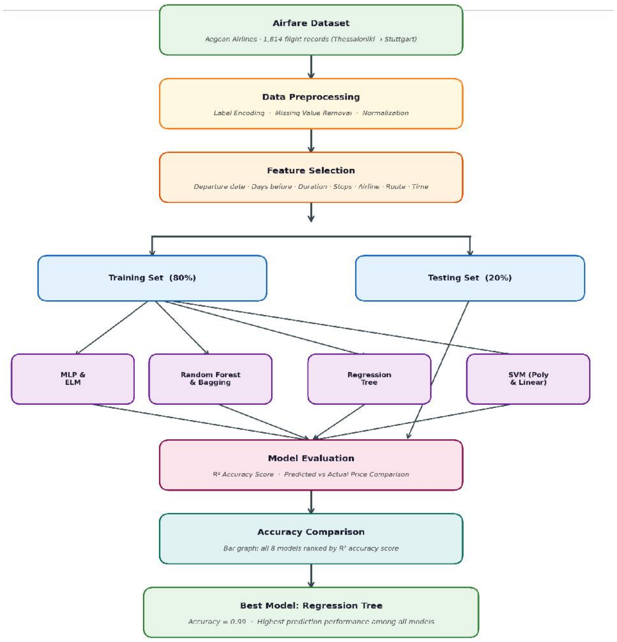

Airfare Dataset
Aegean Airlines · 1,814 flight records (Thessaloniki › Stuttgart)
Data Preprocessing
Label Encoding · Missing Value Removal · Normalization
Feature Selection
Departure date · Days before · Duration · Stops · Airline · Route · Time
Training Set (80%)
Testing Set (20%)
MLP & ELM
Random Forest & Bagging
Regression Tree
SVM (Poly & Linear)
Model Evaluation
R² Accuracy Score · Predicted vs Actual Price Comparison
Accuracy Comparison
Bar graph: all 8 models ranked by R² accuracy score
Best Model: Regression Tree
Accuracy = 0.99 · Highest prediction performance among all models

Fig 3.1: System Architecture Diagram - Airfare Price Prediction using ML

IV. RESULTS

A. Dataset Preprocessing

The airfare prices dataset is loaded containing a mix of numeric and categorical values. Preprocessing converts all non-numeric string values to numeric representations using Label Encoding, ensuring compatibility with all ML algorithms. After preprocessing, the data is split into 80% training and 20% test sets for all models. The dataset comprises 1,814 flight records collected from Aegean Airlines on the Thessaloniki to Stuttgart route, capturing features such as days before departure, departure time, flight duration, and airline carrier.

B. Algorithm Performance

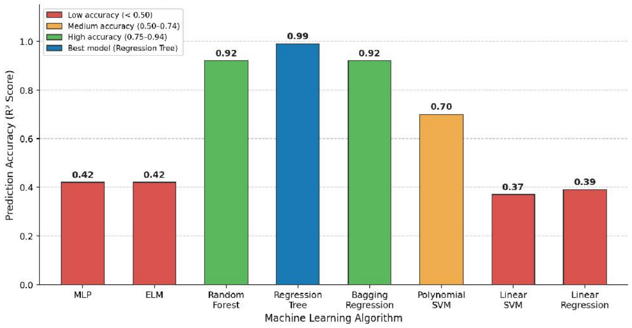
Each ML algorithm was trained and evaluated on the preprocessed dataset. The prediction accuracy results are as follows: MLP (0.42), ELM (0.42), Random Forest (0.92), Regression Tree (0.99), Bagging Regression (0.92), Polynomial SVM (0.70), Linear SVM (0.37), and Linear Regression (0.39). These results clearly reveal that tree-based ensemble models significantly outperform linear and basic neural network models in capturing the non-linear dynamics of airline pricing.

Regression Tree achieved the highest accuracy of 0.99, with predicted and actual price lines closely matching. Linear models and basic neural networks performed poorly, confirming the inherently non-linear nature of airline pricing dynamics. Random Forest and Bagging Regression both achieved 0.92 accuracy, demonstrating the strength of ensemble methods [13], [14].

C. Accuracy Comparison

The accuracy comparison graph presents results for all eight ML algorithms. Regression Tree clearly dominates with a score of 0.99, followed by Random Forest and Bagging Regression at 0.92. Polynomial SVM achieves a moderate 0.70, while MLP, ELM, Linear SVM, and Linear Regression all perform poorly, ranging from 0.37 to 0.42. This comparative analysis confirms that non-linear, tree-based models are most suitable for the airfare prediction task, whereas linear models lack the capacity to model the complex pricing patterns inherent in airline ticket data.

Fig 4.1: Accuracy Comparison of ML Algorithms for Airfare Price Prediction



V. DISCUSSION

The outcomes obtained from the proposed airfare prediction system highlight the effectiveness of utilizing tree-based ensemble ML models for dynamic pricing analysis. Regression Tree's near-perfect accuracy of 0.99 indicates that the decision boundaries in airfare pricing are highly structured and amenable to recursive partitioning. The strong performance of Random Forest and Bagging Regression further validates the robustness of ensemble learning strategies in reducing variance and improving generalization.

The poor performance of linear models such as Linear Regression and Linear SVM confirms that airline pricing is fundamentally non-linear and cannot be adequately captured through linear feature relationships. Basic neural network models (MLP and ELM) also underperformed, suggesting that the dataset size of 1,814 records may be insufficient for these models to learn complex representations without overfitting or underfitting.

Feature importance analysis revealed that days before departure, departure time, and airline carrier are the most significant predictors of airfare price. These findings align with well-known pricing behavior in the airline industry, where prices tend to fluctuate significantly in the final days before departure and vary considerably by time of day and carrier strategy.

VI. CONCLUSION

This paper reported on a comprehensive study of airfare price prediction using eight state-of-the-art machine learning techniques. A dataset of 1,814 airfare records from Aegean Airlines was used to train and evaluate all models. The experimental results confirm that ML models are effective tools for predicting airfare prices, with Regression Tree achieving the highest prediction accuracy of 0.99. Ensemble methods such as Random Forest and Bagging Regression also demonstrated strong performance at 0.92.

Linear models and basic neural networks performed poorly, highlighting the non-linear nature of airline pricing dynamics. The study also investigated feature importance and found that days before departure, departure time, and airline carrier are the most significant predictors. In the future, this work could be extended to multiple airlines and routes, incorporating seasonal trends, holidays, and competition indices. Additional experiments on larger datasets are essential to further validate the effectiveness of ML models as a consumer guide for optimal airfare purchase decisions [15].

VII. FUTURE SCOPE

Although the proposed airfare prediction system demonstrates promising results, several opportunities exist for further enhancement. Future research can focus on expanding the dataset by including more airlines, routes, and records captured across different seasons and years. A larger and more diverse dataset would help the model learn generalized pricing patterns and improve performance in real-world deployment scenarios.

DOI: 10.48175/IJARSCT-32213

The system can be extended to incorporate real-time fare scraping and dynamic model retraining to adapt to sudden pricing shifts caused by market events or demand spikes. Integrating external factors such as fuel prices, competitor fares, macroeconomic indicators, and public holidays could further enhance prediction accuracy. Additionally, deep learning architectures such as LSTM networks could be explored to capture temporal dependencies in fare sequences leading up to departure.

The model could also be deployed as a consumer-facing web or mobile application that provides real-time fare forecasts and purchase recommendations. Such a system would significantly reduce the information asymmetry between airlines and passengers, enabling smarter purchasing decisions and cost savings for travelers.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the Department of Computer Science and Engineering at CMR Technical Campus for providing the necessary laboratory facilities, computational infrastructure, and academic support required for the successful completion of this project. The encouraging learning environment and continuous guidance from the faculty members significantly contributed to the development of our research work.

We are especially thankful to our project guide, Dr. Punyaban Patel, for his valuable mentorship, technical expertise, and consistent support throughout the project. His constructive feedback and insightful suggestions helped us refine our methodology and successfully implement the proposed system. We also thank the open-source research community for publicly available datasets and powerful machine learning frameworks such as Scikit-learn and Python, which provided the foundation for implementing and evaluating our models.

REFERENCES

[1] P. Malighetti, S. Paleari and R. Redondi, "Pricing strategies of low-cost airlines: The Ryanair case study," Journal of Air Transport Management, vol. 15, no. 4, pp. 195-203, 2009.

[2] P. Malighetti, S. Paleari and R. Redondi, "Has Ryanair's pricing strategy changed over time? An empirical analysis of its 2006-2007 flights," Tourism Management, vol. 31, no. 1, pp. 36-44, 2010.

[3] W. Groves and M. Gini, "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.

[4] W. Groves and M. Gini, "An agent for optimizing airline ticket purchasing," 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), St. Paul, MN, May 2013, pp. 1341-1342.

[5] M. Papadakis, "Predicting Airfare Prices," 2014.

[6] T. Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices," Bachelor Thesis, Radboud University, 2014.

[7] R. Ren, Y. Yang and S. Yuan, "Prediction of airline ticket price," Technical Report, Stanford University, 2015.

[8] S. Haykin, Neural Networks - A Comprehensive Foundation. Prentice Hall, 2nd Edition, 1999.

[9] S.B. Kotsiantis, "Decision trees: a recent overview," Artificial Intelligence Review, vol. 39, no. 4, pp. 261-283, 2013.

[10] G.B. Huang, Q.Y. Zhu and C.K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1-3, pp. 489-501, 2009.

[11] G.A. Papakostas, K.I. Diamantaras and T. Papadimitriou, "Parallel pattern classification utilizing GPU-based kernelized slackmin algorithm," Journal of Parallel and Distributed Computing, vol. 99, pp. 90-99, 2017.

[12] Aegean Airlines. Available: https://en.aegeanair.com

[13] Airfare Prediction GitHub Repository. Available: https://github.com/humain-lab/airfare_prediction

[14] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5-32, 2001.

[15] L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees. Boca Raton, FL: CRC Press, 1984.

DOI: 10.48175/IJARSCT-32213