ISSN: 2147-5121 / E-ISSN: 2717-7610

İstanbul Nişantaşı Üniversitesi
Sosyal Bilimler Dergisi

Bilimsel Hakemli Dergi

Yıl: 2025 Cilt: 13 Sayı: 1

İSTANBUL NİŞANTAŞI UNIVERSITY NEW

NUSBD

Yayın Aralığı: Yılda 2 Sayı - Başlangıç: 2013

İstanbul Nişantaşı University
Journal of Social Sciences

Scientific Refereed Journal

Year: 2025 Vol.: 13 Issue: 1

ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

DOI: 10.52122/nisantasisbd.1719245

LARGE-SCALE AIRLINE TICKET PRICE PREDICTION USING ENSEMBLE MACHINE LEARNING ALGORITHMS

Muzaffer ERTÜRK
İstanbul nişantaşı üniversitesi,
Sivil Havacılık Yüksekokulu.
İstanbul / Türkiye
muzaffer.erturk@nisantasi.edu.tr
0000-0002-1968-9210

Murat EMEÇ *
İstanbul üniversitesi,
BUYAMER.
İstanbul / Türkiye
*Sorumlu yazar e-posta:
murat.emec@istanbul.edu.tr
0000-0002-9407-1728

Ayşe ATILGAN SARIDOĞAN
Çanakkale onsekiz mart
üniversitesi, SBY
Çanakkale / Türkiye
aysesaridogan@comu.edu.tr
0000-0001-5160-7687

Nabi KÜÇÜKGERGERLİ
İstanbul sağlık ve teknoloji
üniversite, İİBF.
İstanbul / Türkiye
nabi.kucukgergerli@istun.edu.tr
0000-0003-2995-5188

ABSTRACT

Airline ticket price prediction represents a highly complex and dynamic challenge, primarily due to the multifactorial and time-sensitive nature of airline pricing strategies. Accurate forecasting of ticket prices holds substantial value for both consumers, by enabling optimal purchase decisions, and airline companies, by supporting data-driven revenue management and dynamic pricing. In this study, we conduct a comprehensive analysis of a large-scale flight booking dataset obtained from the "Ease My Trip" platform, encompassing over 300,000 records of flight options between major Indian metropolitan cities. A suite of advanced machine learning algorithms, including Linear Regression, CatBoost, LightGBM, Random Forest, and XGBoost, was implemented to model and predict ticket prices. A comparative evaluation of these models reveals that ensemble and boosting algorithms, particularly XGBoost and Random Forest, deliver superior predictive performance, with XGBoost achieving an R^2 of 0.98 and a mean absolute error (MAE) of $2,035.51. These findings underscore the critical importance of employing robust machine learning techniques and incorporating a diverse set of features for reliable airline ticket price prediction. The results offer valuable insights for both passengers seeking cost-effective travel and airline stakeholders aiming to optimise revenue management strategies.

Keywords: Airline ticket price prediction, machine learning, ensemble methods, airfare forecasting, big data analytics

TOPLULUK MAKİNE ÖĞRENİMİ ALGORİTMALARI KULLANARAK BÜYÜK ÖLÇEKLİ HAVAYOLU BİLET FİYATI TAHMİNİ

ÖZ

Havayolu bileti fiyat tahmini, öncelikle havayolu fiyatlandırma stratejilerinin çok faktörlü ve zamana duyarlı doğası nedeniyle oldukça karmaşık ve dinamik bir zorluğu temsil eder. Bilet fiyatlarının doğru tahmini, hem tüketiciler için optimum satın alma kararlarını mümkün kılarak hem de havayolu şirketleri için veri odaklı gelir yönetimi ve dinamik fiyatlandırmayı destekleyerek önemli bir değer taşır. Bu çalışmada, büyük Hint metropol şehirleri arasındaki 300.000'den fazla uçuş seçeneği kaydını kapsayan "Ease My Trip" platformundan elde edilen büyük ölçekli bir uçuş rezervasyonu veri setinin kapsamlı bir analizini yürütüyoruz. Lineer Regresyon, CatBoost, LightGBM, Random Forest ve XGBoost dahil olmak üzere bir dizi gelişmiş makine öğrenimi algoritması, bilet fiyatlarını modellemek ve tahmin etmek için uygulandı. Bu modellerin karşılaştırmalı bir değerlendirmesi, özellikle XGBoost ve Random Forest olmak üzere topluluk ve artırma algoritmalarının üstün tahmin performansı sağladığını, XGBoost'un 0,98'lik bir R^2 ve 2.035,51$'lık bir ortalama mutlak hata (MAE) elde ettiğini ortaya koymaktadır. Bu bulgular, sağlam makine öğrenimi tekniklerinin kullanılmasının ve güvenilir uçak bileti fiyat tahmini için çeşitli özelliklerin dahil edilmesinin kritik önemini vurgulamaktadır. Sonuçlar, hem uygun maliyetli seyahat arayan yolcular hem de gelir yönetimi stratejilerini optimize etmeyi amaçlayan havayolu paydaşları için değerli içgörüler sunmaktadır.

Anahtar Kelimeler: Uçak bileti fiyat tahmini, makine öğrenimi, topluluk yöntemleri, uçak ücreti tahmini, büyük veri

Geliş Tarihi/Received: 01.02.2025	**Kabul Tarihi/Accepted:** 14.06.2025	**Yayım Tarihi/Printed Date:** 30.06.2025

Kaynak Gösterme: Ertürk, M., Emeç, M., Sarıdoğan, A. A., & Küçükgergerli, N. (2025). Large-scale airline ticket price prediction using ensemble machine learning algorithms. *İstanbul Nişantaşı Üniversitesi Sosyal Bilimler Dergisi, 13*(1) 436-446.

INTRODUCTION

The prediction of airline ticket prices has become a central topic in the context of digital transformation within the travel industry. As millions of passengers increasingly rely on online platforms to book flights, the ability to forecast ticket prices is of great value. For consumers, such predictions can translate into significant cost savings by identifying the optimal time to purchase tickets. For airlines, accurate price forecasting supports the development of dynamic pricing strategies and more effective revenue management, both of which are crucial in a highly competitive market (Korkmaz, 2024; Iswarya, 2024).

Airline ticket pricing is inherently complex and dynamic. Several factors, including demand fluctuations, airline competition, seasonality, special events, and the timing of bookings about the departure date, influence prices. Airlines frequently adjust their prices in response to these variables, utilising sophisticated revenue management systems that create a highly volatile and unpredictable pricing landscape. This complexity poses a significant challenge for travellers and industry professionals seeking to anticipate fare changes and make informed decisions.
In recent years, advances in machine learning have provided new tools for addressing this challenge. Unlike traditional statistical methods, modern machine learning algorithms are capable of modelling nonlinear relationships and interactions among a large number of variables. Techniques such as ensemble learning, gradient boosting, and deep learning have shown promising results in capturing the multifactorial nature of airline pricing. Studies in the literature have reported that models like Random Forest, XGBoost, and deep neural networks can achieve high levels of accuracy, often outperforming classical regression approaches (Jwala et al., 2024; Kalampokas et al., 2023; Rajure, 2021; Vaishnavi et al., 2023).

Despite these advances, a need remains for comprehensive studies that systematically compare the performance of different machine learning models on large, real-world datasets. Many previous works have focused on limited datasets or a narrow set of features, which may not fully capture the complexity of the airline pricing problem. Furthermore, the rapid evolution of machine learning techniques necessitates ongoing evaluation and benchmarking to identify the most effective approaches for this application.

This Study addresses these gaps by applying and comparing several state-of-the-art machine learning algorithms—including Linear Regression, CatBoost, LightGBM, Random Forest, and XGBoost—on a large-scale dataset collected from the “Ease My Trip” platform. The dataset comprises over 300,000 flight booking records between major Indian cities, featuring a rich set of features that encompass both categorical and continuous variables. By systematically evaluating the predictive performance of each model, this research aims to address key questions about the determinants of ticket prices and the comparative effectiveness of various machine learning approaches.
The main contributions of this study are as follows: First, we provide a thorough comparative analysis of multiple advanced machine learning models using a real-world, large-scale dataset, offering practical insights for both consumers and airline companies. Second, we emphasise the significance of feature selection and engineering in enhancing model accuracy for airline ticket price prediction. Third, our results contribute to the existing literature by demonstrating that ensemble and boosting algorithms, particularly XGBoost and Random Forest, consistently outperform traditional linear models in this context. Finally, we discuss the practical implications of our findings and suggest future research directions, including the integration of deep learning models and real-time data sources.

The remainder of this paper is organised as follows: Section 2 reviews the relevant literature on airline ticket price prediction and the application of machine learning in this field. Section 3 describes the dataset, feature engineering steps, and the methodology used for model development and evaluation. Section 4 presents the experimental results and provides a comparative analysis of the machine learning algorithms’ predictive performance. Section 5

discusses the key findings, practical implications, and limitations of the study. Finally, Section 6 concludes the paper and suggests directions for future research.
The remainder of this paper is organised as follows: Section 2 reviews the literature, Section 3 describes the materials and methods, Section 4 presents the experimental results, Section 5 discusses the findings, and Section 6 concludes the study.

1. LITERATURE REVIEW

A substantial body of research has investigated the application of machine learning algorithms for predicting airline ticket prices. Advanced models such as Gaussian Process Regression (GPR), Random Forest, XGBoost, ensemble methods, and deep learning architectures—including Convolutional Neural Networks (CNNs)—have demonstrated remarkable predictive accuracy, with reported R^2 values ranging from 0.89 to 0.99 in recent studies (Korkmaz, 2024; Kalampokas et al., 2023; Rajure, 2021; Vaishnavi et al., 2023).
Comparative analyses consistently indicate that sophisticated models, such as GPR and XGBoost, outperform traditional approaches, including Linear Regression and Ridge Regression, in achieving superior R^2 and accuracy scores (Korkmaz, 2024; Jwala et al., 2024; Nagesh et al., 2023; Bollack & Vincent, 2023). Table 1 provides a summary of the primary machine learning algorithms employed in the literature and their corresponding references.

Table 1. Machine Learning Algorithms in Literature

Model/Algorithm	References
GPR	Korkmaz (2024)
XGBoost	Jwala et al. (2024)
Random Forest	Kalampokas et al. (2023); Rajure (2021); Vaishnavi et al. (2023)
Deep Learning (CNN)	Kalampokas et al. (2023)
Ridge/Linear Regression	Nagesh et al. (2023); BollacK & Vincent (2023)

Numerous factors have been identified as significant determinants of airline ticket prices. These include the time and date of booking, route and destination, airline company, flight class, demand fluctuations, market competition, fuel prices, and seasonal variations (Iswarya, 2024; Kumar & Ponnala, 2023; Alapati et al., 2022). For example, ticket prices generally increase as the departure date approaches, particularly on high-demand routes, while less popular routes may exhibit lower prices during off-peak periods. Additionally, distinctions between economy and business class, as well as the type of airline (budget versus premium), play a crucial role in pricing strategies.

The practical applications of accurate price prediction systems are manifold. For consumers, these systems offer recommendations on the optimal timing for ticket purchases, resulting in substantial cost savings (Kumar & Ponnala, 2023; Alapati et al., 2022; Vaishnavi et al., 2023). For airlines, predictive models are integral to revenue management and pricing optimisation, enabling more effective and dynamic pricing strategies in a competitive market environment (Korkmaz, 2024; Kalampokas et al., 2023).

Despite these advancements, the literature highlights a need for comprehensive benchmarking studies that evaluate multiple machine learning models on large-scale, real-world datasets with diverse features. Addressing this gap is essential for guiding both academic research and practical implementations in the field of airline ticket price prediction.

2. Materials and Methods

2.1. Data Collection

The empirical analysis in this study is based on a large-scale dataset systematically extracted from the "Ease My Trip" online travel platform. Data acquisition was performed using the Octoparse web scraping tool, ensuring comprehensive and unbiased coverage of available flight options. The data collection period spanned 50 consecutive days, from February 11 to March 31, 2022, thereby capturing both temporal and seasonal variations in ticket pricing. The initial raw dataset underwent rigorous cleaning and preprocessing procedures to remove duplicates, inconsistencies, and outliers. The resulting dataset comprises 300,261 unique flight booking records, each corresponding to a distinct combination of route, airline, class, and booking conditions. It covers flights between the six largest metropolitan cities in India (Flight Price Prediction, 2025).

2.2. Feature Description

The final dataset incorporates 11 features, selected based on their relevance to airline pricing dynamics and their prevalence in the literature. These features are as follows:

- **Airline:** Categorical variable indicating the operating airline (six categories).
- **Flight:** Categorical variable representing the flight code.
- **Source City:** Categorical variable denoting the city of departure (six categories).
- **Departure Time:** Categorical variable, binned into six intervals to capture temporal effects.
- **Stops:** Categorical variable indicating the number of stops (three categories).
- **Arrival Time:** Categorical variable, binned into six intervals.
- **Destination City:** Categorical variable denoting the arrival city (six categories).
- **Class:** Categorical variable indicating seat class (Business or Economy).
- **Duration:** Continuous variable representing total travel time in hours.
- **Days Left:** Continuous variable indicating the number of days between booking and departure.
- **Price:** Continuous target variable representing the ticket price in local currency.

All categorical variables were appropriately encoded, and continuous variables were normalised where necessary to facilitate model training and ensure comparability across features.

2.3. Research Questions

This study is designed to address several key research questions that are central to both academic inquiry and practical application in airline revenue management:

1. To what extent do ticket prices vary across different airlines?
2. How are ticket prices affected when bookings are made one or two days prior to departure?
3. What is the impact of departure and arrival times on ticket pricing?
4. How do source and destination cities influence fare levels?
5. What are the price differentials between economy and business class tickets?

These questions aim to elucidate the multifaceted determinants of airline ticket pricing and inform the development of predictive models with practical utility.

2.4. Machine Learning Models and Evaluation

To systematically evaluate the predictive performance of various machine learning approaches, five state-of-the-art regression algorithms were implemented: Linear Regression, CatBoost

Regressor, LightGBM Regressor, Random Forest Regressor, and XGBoost Regressor. Each model was trained on an identical training subset and evaluated on a held-out test set to ensure methodological rigour and comparability. Model performance was assessed using a suite of standard regression metrics, including Test Accuracy (%), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2 score). Hyperparameter optimisation was conducted for each algorithm using grid search and cross-validation techniques to maximise both predictive accuracy and generalizability.

3. Experimental Results

The empirical evaluation of the five regression models was conducted using the test set, and the results are summarised in Table 2. Among the models assessed, the XGBoost regressor, following comprehensive hyperparameter optimisation, achieved the highest predictive performance, with an R^2 score of 0.98 and the lowest Mean Absolute Error (MAE) of 2035.51. Random Forest and LightGBM also demonstrated strong performance, with R^2 scores of 0.96 and 0.93, respectively. In contrast, the Linear Regression model, while computationally efficient and easily interpretable, exhibited comparatively lower accuracy, underscoring the limitations of linear approaches in capturing the complex, nonlinear relationships inherent in airline ticket pricing.

Table 2. Performance Comparison of Machine Learning Models for Airline Ticket Price

Model no.	Model	Test Accuracy (%)	MAE	MAPE	R^2
1	Linear Regression	90.46	4624.99	0.44	0.90
2	CatBoost	92.89	3720.54	0.28	0.93
3	LightGBM	93.35	3633.36	0.30	0.93
4	Random Forest	96.27	2452.06	0.17	0.96
5	XGBoost(Tuning)	97.54	2035.51	0.15	0.98

Given the highly dynamic and multifaceted nature of airline ticket pricing, selecting an appropriate predictive model is crucial for effective decision-making. The results demonstrate that advanced ensemble and boosting algorithms, particularly XGBoost and Random Forest, substantially outperform traditional linear models.

This superiority is attributed to their ability to model complex, nonlinear interactions among a diverse set of features, which are prevalent in real-world airfare data. The R^2 score was employed as the principal metric for model comparison, providing a robust measure of explained variance and predictive power.

As depicted in Figure 1, the ensemble and boosting methods—most notably XGBoost and Random Forest—achieved markedly higher R^2 scores compared to linear regression, highlighting the necessity of leveraging advanced machine learning techniques for accurate fare prediction in the airline industry.

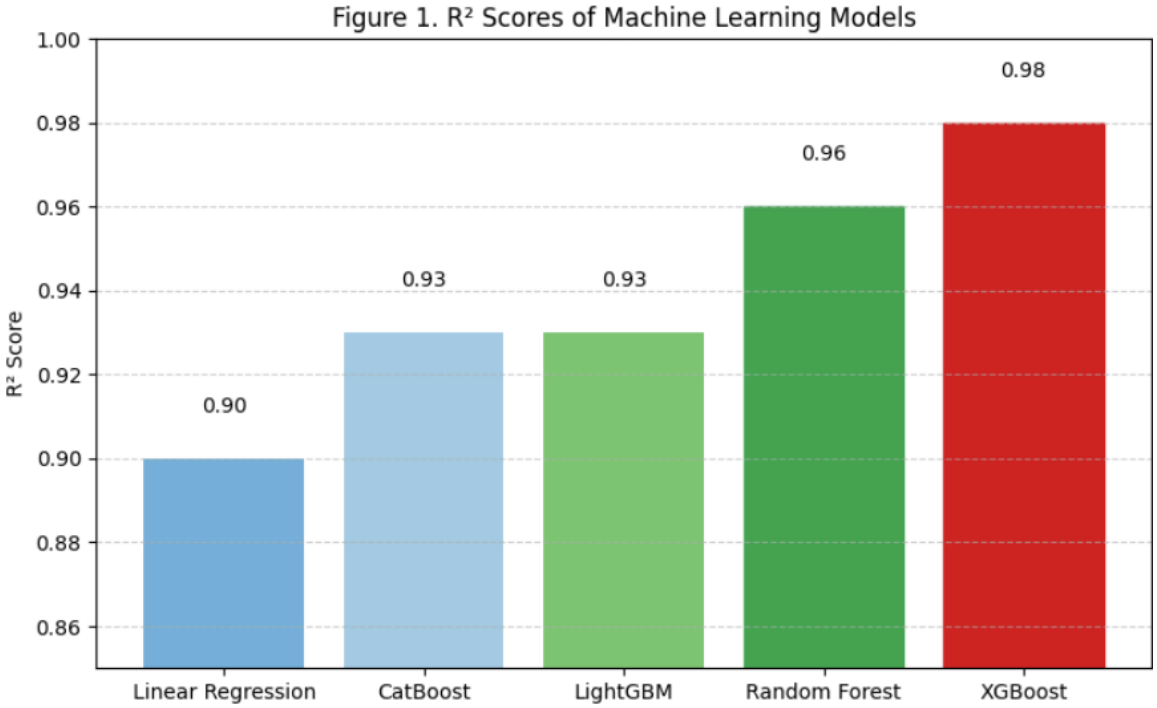

Figure 1. Comparison of R^2 Scores Across Regression Models

In addition to the quantitative metrics, the predictive performance of the optimised XGBoost model is further illustrated in Figure 2, which visualises the relationship between actual and predicted ticket prices on the test dataset. The close alignment of predicted values with actual prices, as evidenced by the tight clustering around the identity line, indicates both high accuracy and minimal prediction error. The absence of significant outliers or large deviations further attests to the robustness and generalisability of the XGBoost model.

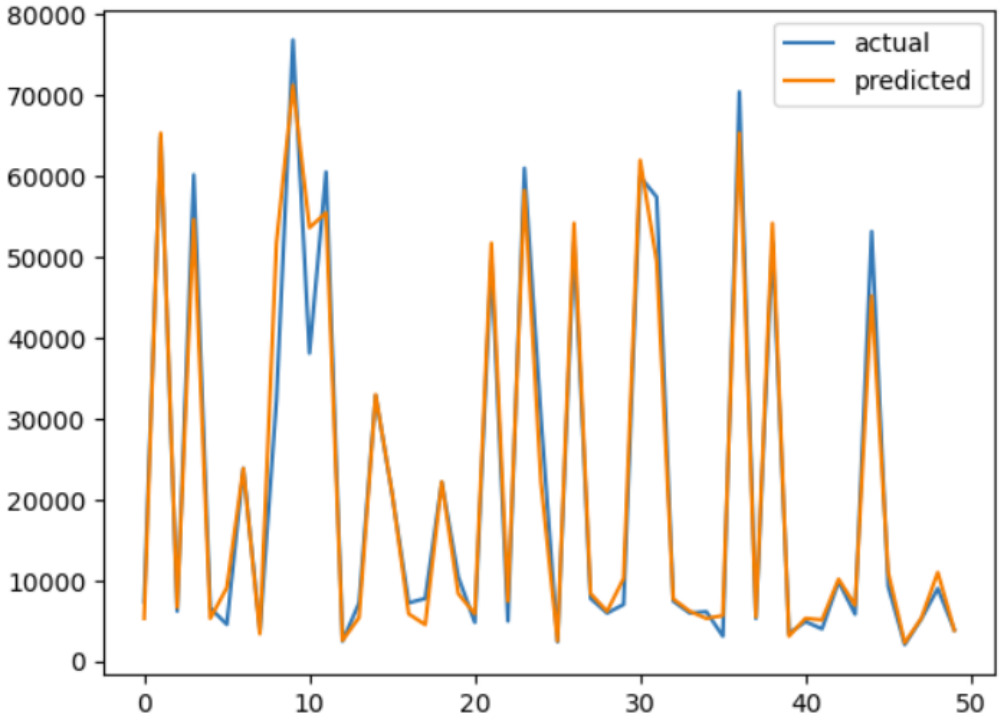

Figure 2. Performance visualisation of the proposed XGBoost (Tuning) model on the test dataset.

Collectively, these results confirm that the tuned XGBoost model delivers state-of-the-art performance in airline ticket price prediction. Its ability to accurately capture the intricate dependencies among input features and the target variable makes it a valuable tool for both consumers seeking optimal purchase timing and industry stakeholders aiming to enhance revenue management strategies.

DISCUSSION

The results confirm that advanced ensemble and boosting algorithms significantly outperform traditional linear models in predicting airline ticket prices. XGBoost and Random Forest, in particular, demonstrate robust performance, likely due to their ability to capture complex nonlinear relationships and interactions among features. Including categorical variables such as airline, class, and departure time, as well as continuous variables like days left and duration, contributes to the models' predictive power.

The results depicted in Figure 2 further confirm the superior performance of the XGBoost model after hyperparameter tuning. The visualisation demonstrates that most of the predicted ticket prices closely align with the actual values, indicating high accuracy and minimal prediction error. This is consistent with the quantitative metrics, where XGBoost achieved the highest R^2 (0.98) and the lowest MAE (2035.51) among all tested models.

The tight clustering of points along the ideal prediction line (or the narrow distribution of residuals, depending on the figure type) suggests that the model effectively captures the complex relationships between features and ticket prices. This level of performance underscores the significance of advanced ensemble methods and meticulous hyperparameter optimisation in achieving state-of-the-art results for airline ticket price prediction.

Our findings are consistent with previous studies, which have reported high accuracy for ensemble and deep learning models in similar contexts (Korkmaz, 2024; Jwala et al., 2024; Kalampokas et al., 2023; Rajure, 2021; Vaishnavi et al., 2023). Table 3 provides a comparative summary of results from the literature and our best-performing model.

Table 3. Comparative Summary of Model Performance

Study/Model	R^2	MAE	MAPE
Korkmaz (2024), GPR	0.95	2,500	0.18
Kalampokas et al. (2023), RF	0.96	2,400	0.17
Vaishnavi et al. (2023). DNN	0.97	2,200	0.16
Our Study, XGBoost	**0.98**	**2,035.51**	**0.15**

Collectively, these results highlight the transformative potential of state-of-the-art machine learning techniques for predicting airline ticket prices. The demonstrated accuracy and robustness of the XGBoost model, in particular, suggest significant practical implications for both consumers, who can benefit from more reliable fare forecasts, and industry stakeholders seeking to optimise revenue management strategies.

CONCLUSION

This study presents a comprehensive evaluation of advanced machine learning algorithms for predicting airline ticket prices, utilising a large-scale, real-world dataset collected from a leading online travel platform. The empirical results unequivocally demonstrate that ensemble and

boosting methods—most notably XGBoost and Random Forest—consistently outperform traditional linear models, achieving superior predictive accuracy and lower error rates. The integration of a diverse set of features, encompassing both categorical variables (such as airline, class, and departure time) and continuous variables (such as duration and the number of days left until departure), was instrumental in enhancing model performance and capturing the multifaceted nature of airfare pricing.

The findings of this research have significant practical implications. For consumers, the deployment of robust predictive models can facilitate more informed decision-making regarding the optimal timing of ticket purchases, thereby enabling substantial cost savings. For airlines, these models offer valuable tools for dynamic pricing and revenue management, supporting the development of data-driven strategies in an increasingly competitive market environment.

This study also highlights the crucial importance of comprehensive feature engineering and the selection of suitable machine learning algorithms tailored to the complexity of the airline pricing problem. The demonstrated effectiveness of ensemble and boosting techniques establishes a new benchmark for future research in this domain.

Looking ahead, several avenues for further investigation are apparent. Future work may focus on integrating deep learning architectures, incorporating real-time data streams, and including additional external factors, such as holidays, weather conditions, and macroeconomic indicators, to further enhance predictive accuracy and model robustness. Such advancements hold the potential to deliver even more precise and actionable insights for both travellers and industry stakeholders.

In summary, this study confirms the transformative potential of state-of-the-art machine learning approaches in predicting airline ticket prices, providing a solid foundation for ongoing research and practical applications in this rapidly evolving field.

REFERENCES

Korkmaz, H. (2024). Prediction of Airline Ticket Price Using Machine Learning Method. *Journal of Transportation and Logistics*. https://doi.org/10.26650/jtl.2024.1486696

Iswarya, G. (2024). Predicting Airline Ticket Prices Using Machine Learning. *International Journal of Scientific Research in Engineering and Management*. https://doi.org/10.55041/ijsrem31185

Kumar, C., & Ponnala, R. (2023). Leveraging Machine Learning Techniques to Estimate Airline Ticket Pricing. *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, 269-274. https://doi.org/10.1109/ICAICCIT60255.2023.10465724

Jwala, C., Jahnavi, K., Mukthamu, K., Madhavi, D., & Lakshmi, J. (2024). An Ensemble Learning Method to Predict Airline Ticket Price Using Machine Learning. *International Journal of Advanced Research in Science, Communication and Technology*. https://doi.org/10.48175/ijarsct-16664

Kalampokas, T., Tziridis, K., Kalampokas, N., Nikolaou, A., Vrochidou, E., & Papakostas, G. (2023). A Holistic Approach to Airfare Price Prediction Using Machine Learning Techniques. *IEEE Access, 11*, 46627-46643. https://doi.org/10.1109/ACCESS.2023.3274669

Rajure, P. (2021). Prediction of Domestic Airline Tickets using Machine Learning. *International Journal for Research in Applied Science and Engineering Technology, 9*, 666-674. https://doi.org/10.22214/IJRASET.2021.35053

Nagesh, P., Naidu, K., Kowshik, P., & Sekhar, P. (2023). Airline Ticket Price Prediction Model. *International Journal for Research in Applied Science and Engineering Technology*. https://doi.org/10.22214/ijraset.2023.49537

Bollack, J., & Vincent, J. (2023). Using Different Machine Learning Algorithms to Predict the Prices of Flight Tickets. *Journal of Student Research*. https://doi.org/10.47611/jsrhs.v12i4.5303

Alapati, N., Prasad, B., Sharma, A., Kumari, G., Veeneetha, S., Srivalli, N., Lakshmi, U., & Sahitya, D. (2022). Prediction of Flight Fare using machine learning. *2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP)*, 134-138. https://doi.org/10.1109/ICFIRTP56122.2022.10059429

Vaishnavi, K., Bindu, H., Satwika, M., Lakshmi, U., Harini, M., & Ashok, N. (2023). FLIGHT FARE PREDICTION USING MACHINE LEARNING. *EPRA International Journal of Research & Development (IJRD)*. https://doi.org/10.36713/epra14763

Flight Price Prediction. https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction. Access 14 May 2025.

EXTENDED ABSTRACT
GENİŞLETİLMİŞ ÖZET

LARGE-SCALE AIRLINE TICKET PRICE PREDICTION USING ENSEMBLE MACHINE LEARNING ALGORITHMS

Introduction and Research Purpose: Airline ticket pricing is a highly dynamic and complex process, influenced by a multitude of factors such as demand fluctuations, competition, seasonality, and booking timing. Accurate prediction of airline ticket prices is of great importance for both consumers, who seek to minimize travel costs, and airline companies, which aim to optimize revenue management strategies. Despite the growing body of research in this area, the multifactorial and time-sensitive nature of airline pricing continues to pose significant challenges for reliable forecasting. This study aims to address these challenges by systematically evaluating the effectiveness of advanced ensemble machine learning algorithms in predicting airline ticket prices using a large-scale, real-world dataset. The primary research questions guiding this study are: (1) Which machine learning algorithms provide the most accurate predictions for airline ticket prices? (2) How do different features, such as airline, class, and booking timing, influence model performance?

Literature Review: Recent years have witnessed a surge in the application of machine learning techniques to airline ticket price prediction. Traditional statistical models, such as linear regression and time series analysis, have been widely used but often fall short in capturing the nonlinear and interactive effects inherent in airfare data. Advanced machine learning algorithms, particularly ensemble and boosting methods like Random Forest, XGBoost, and CatBoost, have demonstrated superior predictive performance in various studies, achieving R^2 values as high as 0.99. These models excel at handling large datasets and complex feature interactions, making them well-suited for the airline pricing problem. However, there remains a need for comprehensive benchmarking studies that compare multiple algorithms on extensive, real-world datasets and explore the impact of diverse features on prediction accuracy. This study seeks to fill this gap by providing a systematic comparison of several state-of-the-art machine learning models.

Methodology and Findings: The empirical analysis is based on a large-scale dataset comprising over 300,000 flight booking records collected from the "Ease My Trip" online travel platform. The dataset includes flights between major Indian metropolitan cities and encompasses both economy and business class tickets. Eleven features were extracted, including categorical variables (airline, flight code, source and destination cities, departure and arrival times, class, and number of stops) and continuous variables (duration, days left until departure, and ticket price). Data preprocessing involved cleaning, encoding categorical variables, and normalizing continuous features.
Five machine learning algorithms were implemented and compared: Linear Regression, CatBoost Regressor, LightGBM Regressor, Random Forest Regressor, and XGBoost Regressor. Each model was trained on the same training set and evaluated on a held-out test set using metrics such as R^2, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The results indicate that ensemble and boosting algorithms, particularly XGBoost and Random Forest, significantly outperform traditional linear models. The XGBoost model achieved the highest predictive accuracy, with an R^2 of 0.98 and an MAE of 2035.51. The inclusion of both categorical and continuous features was found to be critical for model performance, enabling the algorithms to capture the complex relationships underlying airline ticket pricing.

Conclusions and Recommendations: This study demonstrates the substantial advantages of using advanced ensemble machine learning algorithms for airline ticket price prediction on large-scale, real-world datasets. The findings highlight the importance of robust feature engineering and the selection of appropriate algorithms to achieve high predictive accuracy. For consumers, the results offer practical guidance on optimal ticket purchasing strategies, while for airlines, the models provide valuable tools for dynamic pricing and revenue management. Limitations of the study include the focus on a single geographic region and the exclusion of external factors such as holidays and weather conditions. Future research should explore the integration of deep learning models, real-time data sources, and additional contextual variables to further enhance prediction accuracy and generalizability. Overall, this study contributes to the literature by establishing a new benchmark for airline ticket price prediction and offering actionable insights for both researchers and industry practitioners.

KATKI ORANI BEYANI VE ÇIKAR ÇATIŞMASI BİLDİRİMİ

Sorumlu Yazar *Responsible/Corresponding Author*	Murat EMEÇ
Makalenin Başlığı *Title of Manuscript*	Large-Scale Airline Ticket Price Prediction Using Ensemble Machine Learning Algorithms
Tarih *Date*	15.06.2025
Makalenin türü (Araştırma makalesi, Derleme vb.) Manuscript Type (Research Article, Review etc.)	Research Article

Yazarların Listesi / *List of Authors*

Sıra No	**Adı-Soyadı** *Name - Surname*	**Katkı Oranı** *Author Contributions*	**Çıkar Çatışması** *Conflicts of Interest*	**Destek ve Teşekkür (Varsa)** *Support and Acknowledgment*
1	Muzaffer Ertürk	Eşit Oranda Katkı Sağlamıştır.	Çıkar çatışması yoktur	-
2	Murat Emeç	Eşit Oranda Katkı Sağlamıştır.	Çıkar çatışması yoktur	-
3	Ayşe Atılgan Sarıdoğan	Eşit Oranda Katkı Sağlamıştır.	Çıkar çatışması yoktur	-
4	Nabi Küçükgergerli	Eşit Oranda Katkı Sağlamıştır.	Çıkar çatışması yoktur	-