Journal of Transportation and Logistics
**JTL** 2024, 9 (2), 205–218
DOI: 10.26650/JTL.2024.1486696

RESEARCH ARTICLE

# Prediction of Airline Ticket Price Using Machine Learning Method

Hüseyin Korkmaz[1]

[1](Dr.), İstanbul University, Faculty of Transportation and Logistics, İstanbul, Türkiye

**ABSTRACT**
Airline ticket pricing is a complex and dynamic process influenced by various factors, including demand fluctuations, seasonal variations, and competitive strategies. Accurate price prediction is crucial for both airlines, to maximize revenue, and customers, to secure the best deals. Traditional methods often fall short of capturing the intricate and rapidly changing patterns of airfare pricing. With the advent of machine learning algorithms, there is a growing potential to enhance the accuracy and reliability of ticket price predictions. This paper aims to predict ticket prices based on airline flight data using ML algorithms and to compare the performance of ML algorithms. The secondary objective of this paper is to identify the main factors affecting airline ticket prices. The flight and ticket price datasets of THY and PGS that were obtained from open-access sources are used in this paper. The final dataset consists of 962 records for three months from June 1st, 2022 to August 30th, 2022 and includes 19 different variables. Statistical tests and ML algorithms were applied to the final dataset. This paper compares various ML models to predict airline ticket prices, considering performance metrics such as MAE, MSE, RMSE, and R2 during training and test phases. According to the model training and test results, the best algorithm is GPR with R2: 0.86 (training) and R2: 0.90 (test). The findings are consistent with existing literature, further validating the superior efficacy of certain models in specific contexts and demonstrating significant progress in the field. This paper contributes to the literature by comparing the effectiveness of various machine learning algorithms in predicting airline ticket prices, providing new and valuable insights into model performance and key price-determining factors.

**Keywords:** Price Prediction, Ticket Price, Airfare Price, Machine Learning, Intelligent Transportation Systems

## 1. Introduction

In today's highly competitive airline industry, ticket pricing is crucial in attracting customers and maximizing revenue. Airline ticket prices are influenced by a multitude of factors including demand, seasonality, route popularity, fuel costs, and competitor pricing strategies (Wang et al., 2019). Given the dynamic nature of these factors, accurately predicting ticket prices is a complex and challenging task. Traditional pricing models often fall short of capturing the intricate patterns and rapid changes in the market (Deng, 2024).

Recent advancements in Machine Learning (ML) offer promising tools and techniques for addressing this challenge. ML models can analyze vast amounts of historical data and identify hidden patterns that can improve the accuracy of price predictions. These models can be trained to consider a wide range of variables simultaneously, adapting to new data and refining their predictions over time.

This paper aims to explore the application of various ML methods for predicting airline ticket prices. By leveraging ML algorithms, we seek to develop a robust predictive model that can assist airlines in optimizing their pricing strategies. The paper evaluates different ML techniques, including regression models, decision trees, and neural networks, to determine their effectiveness in forecasting ticket prices. The objectives of this research are threefold: (1) to identify the key factors influencing airline ticket prices, (2) to develop and compare the performance of various ML models in predicting these prices, and (3) to provide insights and recommendations for airlines to enhance their pricing strategies using the developed models.

In the following sections, we review the relevant literature on airline pricing and ML applications, describe the methodology and data used in our study, present the results of our model comparisons, and discuss the implications of our findings for the airline industry. This research contributes to the growing body of knowledge on dynamic pricing and offers practical solutions for one of the most critical aspects of airline revenue management.

**Corresponding Author:** Hüseyin Korkmaz **E-mail:** huseyinkorkmaz@istanbul.edu.tr
**Submitted:** 19.05.2024 • **Revision Requested:** 27.06.2024 • **Last Revision Received:** 28.06.2024 • **Accepted:** 10.07.2024 • **Published Online:** 26.07.2024

## 2. Related Work

The prediction of airline ticket prices has been an area of significant research interest, with various approaches and methodologies being explored to enhance accuracy and efficiency. Predicting airline ticket prices is a complex task due to the dynamic nature of pricing influenced by numerous fluctuating factors. Over the past decade, researchers have increasingly employed ML algorithms and Data Mining (DM) techniques to model these prices more accurately (Abdella et al., 2021; Aliberti et al., 2023; Groves & Gini, 2015; Kumar, 2023; Sherly Puspha Annabel et al., 2023; Zhao et al., 2022). This section reviews key studies and their contributions to the field, focusing on the methods employed and their performance outcomes.

Janssen et al. (2014) aimed to predict the lowest ticket price before departure using a linear quantile mixed regression model. While their model demonstrated reasonable short-term performance, its long-term efficiency was found to be inadequate. In this study, multiple LR models were compared to determine the best fit for advising passengers on whether to purchase a ticket immediately or wait for a better price. The authors recommended linear quantile mixed models for predicting the lowest fares, termed "real bargains." This study, however, was limited to economy-class tickets on flights from San Francisco to John F. Kennedy Airport. Tziridis et al. (2017) evaluated eight regression ML models to identify the optimal fare prediction algorithm. Among these, the Bagging Regression model achieved an accuracy of 87.42%, and the Random Forest (RF) Regression Tree achieved 85.91%. On the other hand, Gordiievych and Shubin (2015) utilized the ARIMA model to predict future ticket price drops, although specific performance results were not provided. Another study Santana et al. (2017) proposed Deep Regressor Stacking, which combines RF and Support Vector Machine (SVM) for more accurate predictions, demonstrating the applicability of these techniques across similar domains. Furthermore, Wohlfarth et al. (2011) focused on predicting the best time to buy tickets using Classification And Regression Trees (CART) and RF models, suggesting that these models could offer preliminary advice to customers during pre-registered purchase periods.

Beyond flight-specific features, other factors such as market demand significantly impact ticket pricing. For instance, Huang (2013) used Artificial Neural Network (ANN) and Genetic Algorithm (GA) to predict air ticket sales revenue for a travel agency, incorporating variables like international oil prices and stock market indices. The GA optimized input features for the ANNs, resulting in a Mean Absolute Percentage Error (MAPE) of 9.11%. Another study Kalampokas et al. (2023) examines airfare price prediction by comparing the pricing policies of different airlines using AI techniques. Specifically, it extracts features from over 136,000 flights from Aegean, Turkish, Austrian, and Lufthansa Airlines across six popular international destinations. AI models from three domains—ML, DL, and Quantum ML (QML)—comprising 16 different architectures, were employed to predict ticket prices. The findings indicate that at least three models from each domain achieved accuracy rates between 89% and 99% for this regression problem, demonstrating the effectiveness of AI in airfare price prediction.

Early research explored classification models to predict price trends. For instance, Ren et al. (2014) developed an ensemble model incorporating Linear Regression (LR), Naive Bayes, Softmax Regression, and SVM to predict the lowest ticket price before departure. The training errors for Naive Bayes and Softmax Regression were reduced to 24.88% and 20.22%, respectively, with SVM also showing an approximate 1% reduction in error. However, their SVM regression model underperformed, leading them to use an SVM classification model to differentiate prices as "higher" or "lower" than the average.

Gui et al. (2020) applied an ensemble model combining random tree models with Deep Learning (DL) to predict flight delays. The Long Short-Term Memory (LSTM) network effectively handled aviation sequence data, while the RF model achieved a 90.2% accuracy for binary classification, thereby mitigating overfitting issues. Likewise, Shih et al. (2019) proposed a DL model utilizing an attention mechanism for multivariate time series prediction, which incorporated frequency domain information for forecasting. Their TPA-LSTM model outperformed others in experimental tests, demonstrating the effectiveness of integrating attention mechanisms with DL techniques for improving prediction accuracy.

Lai et al. (2018) explored multivariate time series prediction using DL models, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), to extract short-term local dependency patterns among variables. Their results indicated superior performance across three out of four experimental datasets. Similarly, Yujing et al. (2020) focused on predicting flight passenger load factors with a CNN model incorporating a multi-granularity temporal attention mechanism (MTA-RNN), which demonstrated the best performance in their experiments. In the realm of stock time series prediction, Qin et al. (2017) applied a dual-stage attention-based RNN (DA-RNN), which showed superior performance on the SML 2010 and NASDAQ 100 datasets compared to other models. Additionally, Chen et al. (2018) developed a dual-stage attention-based RNN for sales volume forecasting in a commercial scenario, integrating trend alignment with dual-attention, multi-task RNNs, and their trend alignment with dual-attention model yielded the best prediction results in their studies. Collectively, these studies highlight the effectiveness of advanced DL techniques in various predictive tasks, demonstrating their potential for superior performance across different domains.

The collection and processing of airline ticket data, often sourced from web crawling or private collaborations, pose significant challenges. This variability makes it difficult to replicate studies and compare model performances. On the other hand, these studies illustrate the breadth of approaches and techniques applied to airline ticket price prediction, from traditional regression models to advanced ML and DL frameworks. Each method has its strengths and limitations, contributing to the ongoing development of more accurate and efficient prediction models in this dynamic field.

## 3. Method

This paper aims to predict ticket prices based on airline flight data using ML algorithms and to compare the performance of ML algorithms. The secondary objective of this paper is to identify the main factors affecting airline ticket prices. The flight and ticket price datasets from Turkish Airlines (THY) and Pegasus Airlines (PGS) are used in this paper. The data obtained were analyzed using DM. DM allows the study data to be analyzed accurately and reduces the error rate. Thus, time and performance losses are prevented. Statistical tests and ML algorithms were applied to the final extracted dataset and airline ticket price prediction was performed.

### 3.1. Data Preprocessing

The dataset used for the analysis in this paper was obtained from open-access sources. The final dataset consists of 962 records for 3 months from June 1st, 2022 to August 30th, 2022. The dataset includes flight and ticket price data of THY and PGS airlines. There are 19 different variables in the dataset.

In this paper, some sub-processes were performed within the scope of data preprocessing. These are: removing columns with out-of-scope and missing values, completing missing data, extracting outliers and repeated values, data editing, normalization, and standardization. In general, identifying missing data, extracting outliers, and cleaning the dataset improves the accuracy of the models and makes the results more reliable.

Data editing, in general, can be expressed as cleaning categorical variables and re-expressing them with a standard. The data editing phase was completed with two approaches which are called one-hot encoding and label encoding. Data editing can be defined as the rescaling and standardization of numerical variables in general. First, the data were arranged, and then normalization and standardization were performed. On the other hand, the categorical variables were transformed to numerical classification, labelled, and categorized by removing the textual expressions in the dataset.

Normalization and standardization are two important techniques used in data preprocessing. Normalization rescales the data between 0 and 1, while standardization rescales the data to have the same mean (0) and the same standard deviation (1) (Karataş, 2021). The min-max scaler method which is one of the most popular normalization methods was used in this paper. Notations of normalization and standardization are presented in Table 1.

**Table 1.** Notations of normalization and standardization

| | Notation | Explanation |
|---|---|---|
| **Normalization** | $X_{new} = \dfrac{X - X_{min}}{X_{max} - X_{min}}$ | $X_{new}$ : new normalized value<br>$X_{max}$ : maximum value in variable<br>$X_{min}$ : minimum value in variable |
| **Standardization** | $Z = \dfrac{X - \mu}{S}$ | $Z$ : new standardized z-score<br>$X$ : number of observations<br>$\mu$ : mean of observations<br>$S$ : standard deviation |

### 3.2. Data Analysis

Some statistical analyses were performed on the final dataset. Statistical analysis is a type of analysis using statistical methods to make sense of the data, identify patterns, understand the relationships between variables, and predict future events. In this paper, descriptive analysis and correlation analysis were performed within the scope of statistical analysis.

Descriptive analysis is used to understand the general characteristics of the dataset, to identify important patterns and trends, and to provide a basis for further analysis. Descriptive analysis aims to reveal the general characteristics of the variables related to flight and ticket price in the final dataset in detail. The study dataset and the results of the descriptive analysis are shown in Table 2.

**Table 2.** Study dataset and descriptive analysis results

| Variable Name | Description | Data Type | Unique Value Count | Min. | Max. | Range | Mean | Median | Mode | Standard Deviation | Variance |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AirlineCode | The commercial code of the airline | Categorical | 2 | 1 | 2 | 1 | 1,25 | 1 | 1 | - | - |
| FlightDay | The day of flight | Numerical | 30 | 1 | 31 | 30 | 12,77 | 11 | 1 | 8,96 | 80,35 |
| FlightMonth | The month of flight | Numerical | 3 | 6 | 8 | 2 | 6,99 | 7 | 6 | 0,82 | 0,68 |
| FlightYear | The year of flight | Numerical | 1 | 2022 | 2022 | 0 | 2022 | 2022 | 2022 | 0 | 0 |
| DepCode | The commercial code of departure airport | Categorical | 29 | 1 | 29 | 28 | 12,83 | 12,5 | 1 | - | - |
| DesCode | The commercial code of arrival airport | Categorical | 9 | 21 | 37 | 16 | 31,62 | 33 | 21 | - | - |
| DepHour | The departure time in hours | Numerical | 22 | 0 | 23 | 23 | 10,44 | 10 | 11 | 5,17 | 26,69 |
| DepMin | The departure time in minutes | Numerical | 12 | 0 | 55 | 55 | 23,57 | 20 | 0 | 17,79 | 316,35 |
| ArrHour | The arrival time in hours | Numerical | 24 | 0 | 23 | 23 | 12,93 | 14 | 7 | 6,15 | 37,79 |
| ArrMin | The arrival time in minutes | Numerical | 12 | 0 | 55 | 55 | 26,37 | 25 | 30 | 17,62 | 310,55 |
| Day | Whether the arrival was the same day or next day | Categorical | 2 | 1 | 2 | 1 | 1,11 | 1 | 1 | - | - |
| Duration(min) | The flight duration | Numerical | 132 | 55 | 1465 | 1410 | 314,96 | 280 | 230 | 184,72 | 34120,63 |
| BoughtDay | The day that the ticket was bought | Numerical | 3 | 16 | 18 | 2 | 16,54 | 16 | 16 | 0,75 | 0,56 |
| BoughtMonth | The month that the ticket was bought | Numerical | 1 | 2 | 2 | 0 | 2 | 2 | 2 | 0 | 0 |
| BoughtYear | The year that the ticket was bought | Numerical | 1 | 2022 | 2022 | 0 | 2022 | 2022 | 2022 | 0 | 0 |
| Stop | The number of flight stops | Numerical | 3 | 1 | 3 | 2 | 1,85 | 2 | 2 | 0,49 | 0,24 |
| DayDiffer | The day difference between date of flight and ticket sale | Numerical | 82 | 103 | 194 | 91 | 146,38 | 144 | 136 | 27,12 | 735,73 |
| PriceEx | The price exchange between date of flight and ticket sale | Numerical | 20 | 0 | 99 | 99 | 74,46 | 81 | 98 | 28,26 | 798,72 |
| Price(₺) | The price of airline ticket | Numerical | 148 | 279 | 1212 | 933 | 935,85 | 1005 | 1159 | 250,33 | 62663,68 |

Correlation analysis is a statistical technique that measures the relationship between two or more variables and the strength of this relationship. Correlation analysis is usually performed using the Pearson Correlation Coefficient (Miles & Banyard, 2007). Pearson correlation is a parametric measure of the linear relationship between two continuous variables (Gibbons, 1997; Howell, 1992). Its values vary between -1 and +1 (Cohen, 2013). +1 is a full positive correlation, meaning that when one variable increases, the other variable increases linearly. 0 means that there is no correlation between the two variables. -1 means full negative correlation, meaning that when one variable increases, the other variable decreases linearly. According to the results of the correlation analyses, the variables with a correlation between them and the related values are presented in Table 3.

**Table 3.** Correlation analysis results

| | AirlineCode | FlightDay | FlightMonth | FlightYear | DepCode | DesCode | DepHour | DepMin | ArrHour | ArrMin | Day | Duration(min) | BoughtDay | BoughtMonth | BoughtYear | Stop | DayDiffer | PriceEx | Price(₺) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AirlineCode** | **1,00** | | | | | | | | | | | | | | | | | | |
| **FlightDay** | 0,09 | **1,00** | | | | | | | | | | | | | | | | | |
| **FlightMonth** | -0,03 | 0,04 | **1,00** | | | | | | | | | | | | | | | | |
| **FlightYear** | 0,00 | 0,00 | 0,00 | **1,00** | | | | | | | | | | | | | | | |
| **DepCode** | 0,06 | -0,12 | -0,04 | 0,00 | **1,00** | | | | | | | | | | | | | | |
| **DesCode** | 0,02 | 0,03 | 0,00 | 0,00 | -0,04 | **1,00** | | | | | | | | | | | | | |
| **DepHour** | 0,22 | -0,02 | 0,03 | 0,00 | 0,10 | 0,02 | **1,00** | | | | | | | | | | | | |
| **DepMin** | -0,15 | 0,17 | 0,04 | 0,00 | -0,18 | 0,00 | -0,13 | **1,00** | | | | | | | | | | | |
| **ArrHour** | -0,06 | 0,02 | -0,04 | 0,00 | 0,15 | 0,01 | 0,23 | -0,21 | **1,00** | | | | | | | | | | |
| **ArrMin** | -0,02 | 0,02 | 0,02 | 0,00 | -0,02 | -0,09 | -0,01 | -0,03 | -0,08 | **1,00** | | | | | | | | | |
| **Day** | 0,18 | -0,06 | 0,05 | 0,00 | 0,07 | -0,03 | 0,53 | 0,08 | -0,61 | 0,06 | **1,00** | | | | | | | | |
| **Duration(min)** | -0,04 | -0,09 | -0,01 | 0,00 | 0,33 | -0,08 | 0,11 | -0,10 | 0,13 | 0,09 | 0,37 | **1,00** | | | | | | | |
| **BoughtDay** | 0,43 | -0,28 | -0,07 | 0,00 | 0,58 | -0,03 | 0,27 | -0,25 | 0,06 | -0,03 | 0,26 | 0,33 | **1,00** | | | | | | |
| **BoughtMonth** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | **1,00** | | | | | |
| **BoughtYear** | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | **1,00** | | | | |
| **Stop** | -0,14 | -0,07 | 0,00 | 0,00 | 0,24 | -0,18 | -0,05 | -0,04 | 0,15 | 0,05 | 0,12 | 0,69 | 0,13 | 0,00 | 0,00 | **1,00** | | | |
| **DayDiffer** | -0,01 | 0,38 | 0,94 | 0,00 | -0,09 | 0,01 | 0,01 | 0,10 | -0,04 | 0,03 | 0,02 | -0,04 | -0,18 | 0,00 | 0,00 | -0,03 | **1,00** | | |
| **PriceEx** | 0,49 | 0,21 | 0,01 | 0,00 | -0,15 | 0,05 | 0,13 | 0,08 | -0,07 | -0,02 | 0,06 | -0,22 | 0,04 | 0,00 | 0,00 | -0,35 | 0,08 | **1,00** | |
| **Price(₺)** | -0,29 | 0,05 | 0,09 | 0,00 | 0,22 | -0,09 | -0,07 | 0,02 | 0,22 | 0,08 | 0,00 | 0,56 | -0,02 | 0,00 | 0,00 | 0,80 | 0,10 | -0,36 | **1,00** |

Correlation analysis was applied to the final dataset variables used in this paper. According to the correlation matrix presented in Table 3, correlations less than -0.50 and greater than 0.50 are highlighted in red. According to the analysis results, it was found that there were high positive correlations between some variables. There is a positive linear relationship of 0.94 between the DayDiffer and FlightMonth variables and 0.80 between the Price(£) and Stop variables. In this case, the Multicollinearity problem arises. Multicollinearity refers to a situation where there is a high correlation or relationship between independent variables in a predicting model (Farrar & Glauber, 1967). In other words, one or more independent variables are strongly correlated with other independent variable(s). This may cause problems in the prediction model. At this point, the variables the FlightMonth and Stop which cause positive correlation and are also considered to negatively affect the model performance were removed from the model to perform the prediction accurately.

### 3.3. Modelling

This paper requires a supervised regression ML technique according to the dataset structure and the problem addressed. Regression algorithms, which are used in cases where the dependent variable should take a continuous value, stand out with their data analysis and prediction capabilities. In this study, regression algorithms are used within the scope of supervised learning on a labelled dataset. In this paper, various regression algorithms such as Linear Regression (LR), Decision Tree (DT), Support Vector Machines (SVM), Efficiently Linear Regression (ELR), Gaussian Process Regression (GPR), Kernel Approximation Regression (KAR), Ensemble Tree (ET), and Neural Networks (NN), which are widely used in the literature, were used and their performances were compared.

LR is a simple, yet powerful statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fitting straight line (or hyperplane in higher dimensions) through the data points. DT is a type of decision tree used for regression tasks. They recursively partition the feature space into smaller regions and fit a simple model (usually a constant value) to each region. This allows for non-linear relationships to be captured in the data. SVM is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates the data into different classes while maximizing the margin. SVM can also be used for regression tasks by finding the hyperplane that best fits the data within a margin of tolerance. ELR is an optimization approach that combines stochastic gradient descent with L1 and L2 regularization to efficiently solve linear regression problems. It aims to minimize the sum of squared errors while penalizing large coefficients to prevent overfitting. GPR is a non-parametric probabilistic approach to regression tasks. It models the relationship between input and output variables as a joint Gaussian distribution, allowing for uncertainty estimation in predictions. GPR is flexible and can capture complex relationships without assuming a specific parametric form. KAR is a regression method that approximates the kernel trick used in Support Vector Machines. It maps input features into a higher-dimensional space using a kernel function, allowing linear models to capture non-linear relationships efficiently. ET, commonly known as Random Forests or Gradient-Boosted Trees, are ensemble learning techniques that combine multiple decision trees to improve predictive performance. They work by training multiple trees independently and averaging their predictions (or combining them in a weighted manner) to make more accurate predictions. NN is a class of ML models inspired by the structure and function of the human brain. They consist of interconnected nodes organized in layers and are capable of learning complex patterns and relationships from data. Neural networks have been successful in various tasks, including regression, classification, and pattern recognition.

LR serves as the fundamental basis for many regression techniques, including ELR, which enhances LR's optimization process. SVM builds upon LR's principles, providing a robust framework for both classification and regression tasks, with ELR borrowing optimization strategies from SVM. DT offers a different approach, utilizing tree-based structures to capture non-linear relationships, close to ET which leverages multiple decision trees for improved accuracy. GPR diverges by employing a probabilistic framework, allowing for uncertainty estimation in predictions, contrasting KAR which efficiently approximates non-linear relationships using kernel functions, reminiscent of SVM's kernel trick. NN stands as a versatile paradigm, capable of learning intricate patterns and relationships from data, with the potential to encompass aspects of all aforementioned methods within their deep architectures. The regression algorithms used in this paper and their notations are presented in Table 4.

**Table 4.** Regression algorithms and notations

| Algorithm | Notation | Explanation |
|---|---|---|
| LR | $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$ | $y$ : dependent variable<br>$x_n$ : independent variable<br>$\beta_n$ : regression coefficient<br>$\epsilon$ : error term |
| DT | $\hat{y} = f(x) = \sum_{m=1}^{M} c_m \, . \, I(x \in R_m)$ | $\hat{y}$ : predicted value of dependent variable<br>$x$ : vector of dependent variable<br>$M$ : total number of leaf nodes<br>$R_m$ : region corresponding to the *m*-th leaf node<br>$c_m$ : prediction coefficient at the *m*-th leaf node<br>$I(x \in R_m)$ : function indicating whether $x$ is present in $R_m$ |
| SVM | $\rightarrow w \, . \, x + b = 0$ (hyperplane)<br>$\rightarrow min_{w,b} \frac{1}{2} \|w\|^2$<br>$\rightarrow y_i(w \, . \, x_i + b) \geq 1 \; for \; i = 1,2,\ldots,n$ (constraints) | $w$ : weight vector perpendicular to the hyperplane<br>$b$ : bias term<br>$x$ : input feature vector<br>$x_i$ : feature vector<br>$y_i$ : class label |
| ELR | $J(w) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_w(x^{(i)}) - y^{(i)}\right)^2 + \lambda_1 \|w\|_1 + \frac{\lambda_2}{2} \|w\|_2^2$ | $w$ : weight vector<br>$m$ : number of training samples<br>$h_w(x^{(i)})$: predicted value for the *i*-th observation<br>$y^{(i)}$ : actual value for the *i*-th observation<br>$\lambda_1$ and $\lambda_2$: L1 and L2 regularization parameters |
| GPR | $f(x) \sim GP(m(x), K(x, x'))$ | $f(x)$ : predicted value of dependent variable<br>$m(x)$ : mean function<br>$K(x, x')$ : kernel function |
| KAR | $f(x) = \sum_{i=1}^{n} a_i \, . \, K(x, x_i)$ | $f(x)$ : predicted value of dependent variable<br>$x$ : input value to be estimated<br>$x_i$ : *i*-th observation in the dataset<br>$a_i$ : coefficient of *i*-th observation<br>$K(x, x_i)$ : kernel function |
| ET | $y(x) = \frac{1}{B} \sum_{i=1}^{B} T_i(x)$ | $y(x)$ : predicted output<br>$B$ : number of trees<br>$T_i(x)$ : prediction of the *i*-th tree for $x$ |
| NN | $y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$ | $y$ : output of a neural cell<br>f : activation function (i.e. sigmoid, ReLU, tanh)<br>$x_i$ : input value<br>$w_i$ : input weight<br>$b$ : bias term |

### 3.4. Performance Measures

R-squared - Determination Coefficient ($R^2$), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) metrics were used to measure the performance of the ML algorithms in this paper. $R^2$ indicates that the independent variable can explain the percentage of total changes in the dependent variable. $R^2$, which ranges between 0 and 1, indicates that the model performs better as it approaches 1. MSE is another important metric that evaluates the model performance. MSE measures the mean squared error between the actual and predicted values of the model. Lower MSE values indicate that the model makes better predictions. RMSE is the square root of MSE. Calculating the RMSE is a useful way to evaluate the model's accuracy. RMSE measures the error between the actual values and the predicted values. The error rate decreases as the RMSE approaches 0. The performance measures used in this paper and their notations are presented in Table 5.

**Table 5.** Performance measures and notations

| Performance Measures | Explanation | Reference |
|---|---|---|
| $R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$ | $n$ : number of observations<br>$y_i$ : true responses<br>$\hat{y}_i$ : predicted responses<br>$\bar{y}_i$ : true responses mean | (Barrett, 2000; Di Bucchianico, 2008) |
| $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ | $n$ : number of observations<br>$y_i$ : true responses<br>$\hat{y}_i$ : predicted responses | (Hyndman & Koehler, 2006; Makridakis et al., 1982) |
| $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$ | $n$ : number of observations<br>$y_i$ : true responses<br>$\hat{y}_i$ : predicted responses | (Hyndman & Koehler, 2006; Nevitt & Hancock, 2000) |
| $MAE = \frac{1}{n} \sum_{i=1}^{n} \|y_i - \hat{y}_i\|$ | $n$ : number of observations<br>$y_i$ : true responses<br>$\hat{y}_i$ : predicted responses | (Hyndman & Koehler, 2006; Sammut & Webb, 2010) |

## 4. Results

In the results section, it is focused on the performance results of the implemented regression-based ML algorithms. MATLAB R2023b was used in the training and test processes of ML algorithms. MATLAB is a widely preferred tool in data analysis and ML applications and played a crucial role in this study. The performance results provide a detailed perspective on the effectiveness, accuracy, and reliability of the models developed to predict airline ticket prices. The findings will provide an important basis for how successful these models can be in practical applications.

According to the model training and test results, the best algorithm is GPR with $R^2$: 0.86 (training) and $R^2$: 0.90 (test). The GPR algorithm has lower RMSE, MSE, and higher $R^2$ values, indicating that this model has a better learning and prediction capability. However, it should be noted that each model may have various advantages and disadvantages in different application and problem contexts. The choice of the best model may vary depending on the requirements and data structure of a particular problem. Figure 1 shows the airline ticket prices and their predicted values from the final dataset. Also, Figure 2 shows the comparison of the observed Price(TL) values in the training data set with the predicted values according to the perfect prediction curve (diagonal).

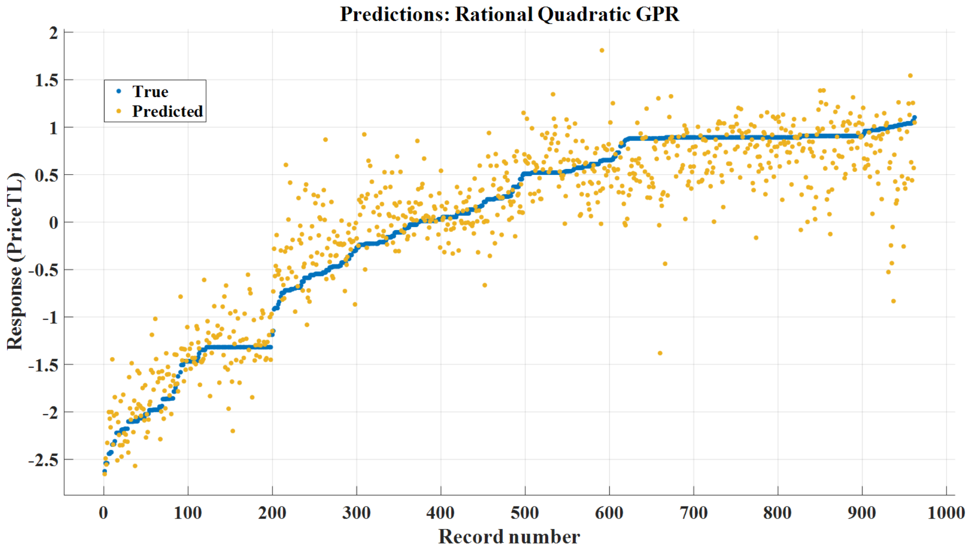


**Figure 1.** Distribution of true and predicted values by GPR algorithm (Training)

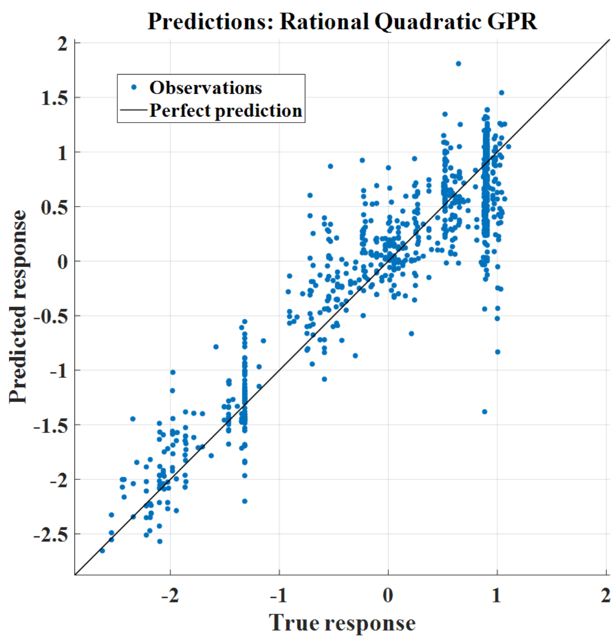


**Figure 2.** Comparison of true and predicted values by GPR algorithm (Training)

### 4.1. Training and Test Dataset

The numbers and percentages of the datasets used in the training and testing processes of the implemented ML algorithms are presented in this part. The results obtained are critical for the evaluation of the performance. The final dataset used in this paper is divided into a 90% training set and a 10% test set. The number and proportions of training and test datasets used for ML algorithms are presented in Table 6.

**Table 6.** The count and percentage of observations for training and test datasets

| | Observations | Percentage |
|---|---|---|
| **Training Data** | 866 | %90 |
| **Test Data** | 96 | %10 |
| **Total** | 962 | %100 |

### 4.2. Performance Optimization

In this paper, different methods are used for the performance optimization of regression-based ML models. These methods are k-fold cross validation, principal component analysis (PCA), and feature selection (FS). A value of 5 was set for k-fold cross validation. 5-fold cross validation is a technique that involves dividing the dataset into five equal parts. One of these parts is used as the test dataset, while the remaining four parts are used as the training dataset. The model is then trained and tested, and this process is repeated five times, each time choosing a different part as the test dataset. The overall model performance is calculated by averaging the performance measures obtained from each test run. PCA analysis was performed as another method. PCA results can be used to assess the significance and predictive power of variables in regression models.

The FS algorithm was implemented to reduce the model complexity, avoid overlearning, and increase the predictive power. As a result of the prediction model experiments, it has been observed that FS analyses affect the prediction performance by about 3-5%. The variables selected, the tests performed and the scores of the variables within the scope of the FS analysis are presented in Table 7. In this paper, F Test scores were used. Accordingly, 13 different variables with scores greater than zero were selected to be used in the training and testing of the models.

**Table 7.** Feature selection algorithms and importance scores of variables

| No. | Features | F Test | No. | Features | MRMR |
|---|---|---|---|---|---|
| 1 | Duration(min) | 5.485.005 | 1 | PriceEx | 16.096 |
| 2 | PriceEx | 3.893.397 | 2 | DayDiffer | 10.713 |
| 3 | DesCode | 2.984.395 | 3 | DesCode | 0.8917 |
| 4 | DepCode | 483.303 | 4 | Duration(min) | 0.7631 |
| 5 | BoughtDay | 480.109 | 5 | ArrHour | 0.6174 |
| 6 | ArrHour | 459.737 | 6 | BoughtDay | 0.5873 |
| 7 | AirlineCode | 441.948 | 7 | DepMin | 4416 |
| 8 | DepHour | 353.904 | 8 | FlightDay | 0.3541 |
| 9 | FlightDay | 134.514 | 9 | ArrMin | 0.3432 |
| 10 | ArrMin | 108.690 | 10 | DepHour | 0.2586 |
| 11 | DayDiffer | 48.172 | 11 | AirlineCode | 0.1939 |
| 12 | DepMin | 40.774 | 12 | DepCode | 0.1725 |
| 13 | Day | 0.0362 | 13 | Day | 0.0258 |

### 4.3. Model Performance Comparison

The results of our experiments aimed at comparing the performance of various ML algorithms are presented in this section. Below, detailed information is provided on the performance error rates (MAE, MSE, RMSE) and $R^2$ values for each model type on both training and test sets. Additionally, the computation times of the models are compared. This analysis allows us to evaluate the performance of the models in terms of accuracy and computational cost. Comparative training and test results of regression-based ML algorithms are presented in Table 8.

**Table 8.** Comparative training and testing results

| | | Training | | | | | Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Model Category** | **Model Type** | **MAE** | **MSE** | **RMSE** | **$R^2$** | **Time (obs/sec)** | **MAE** | **MSE** | **RMSE** | **$R^2$** |
| LR | Linear | 0,40039 | 0,29095 | 0,5394 | 0,70929 | 3600 | 0,38031 | 0,27258 | 0,52209 | 0,72375 |
| | Interactions Linear | 33,74 | 87300 | 295,47 | -87227 | 2000 | 0,70971 | 4,2857 | 2,0702 | -3,3435 |
| | Robust Linear | 0,36837 | 0,33751 | 0,58095 | 0,66277 | 3700 | 0,39339 | 0,38422 | 0,61986 | 0,6106 |
| | Stepwise Linear | 0,54087 | 1,392 | 1,1798 | -0,39087 | 3500 | 0,42473 | 0,83185 | 0,91206 | 0,15692 |
| DT | Fine Tree | 0,33102 | 0,31785 | 0,56379 | 0,68241 | 4600 | 0,29416 | 0,29696 | 0,54494 | 0,69904 |
| | Medium Tree | 0,34308 | 0,28885 | 0,53744 | 0,71139 | 5300 | 0,36839 | 0,3555 | 0,59624 | 0,63971 |
| | Coarse Tree | 0,3962 | 0,32469 | 0,56981 | 0,67558 | 7200 | 0,42497 | 0,38576 | 0,62110 | 0,60903 |
| SVM | Linear SVM | 0,37302 | 0,30796 | 0,55494 | 0,69229 | 6600 | 0,38296 | 0,31851 | 0,56437 | 0,67719 |
| | Quadratic SVM | 0,31177 | 0,21702 | 0,46585 | 0,78316 | 5200 | 0,29941 | 0,1584 | 0,398 | 0,83946 |
| | Cubic SVM | 0,29949 | 0,17651 | 0,42013 | 0,82364 | 6700 | 0,23893 | 0,10542 | 0,32468 | 0,89316 |
| | Fine Gaussian SVM | 0,76674 | 0,88528 | 0,94089 | 0,11545 | 6800 | 0,76768 | 0,8666 | 0,93091 | 0,12171 |
| | Medium Gaussian SVM | 0,3146 | 0,18329 | 0,42813 | 0,81686 | 6400 | 0,25806 | 0,11261 | 0,33558 | 0,88587 |
| | Coarse Gaussian SVM | 0,4651 | 0,33409 | 0,578 | 0,66619 | 4200 | 0,41318 | 0,25137 | 0,50137 | 0,74524 |
| ELR | Efficient Linear Least Squares | 0,40347 | 0,29215 | 0,54051 | 0,70809 | 3400 | 0,38285 | 0,27045 | 0,52005 | 0,7259 |
| | Efficient Linear SVM | 0,36477 | 0,31808 | 0,56399 | 0,68218 | 5200 | 0,3778 | 0,33579 | 0,57948 | 0,65968 |
| ET | Boosted Trees | 0,33362 | 0,21426 | 0,46288 | 0,78592 | 3400 | 0,33647 | 0,23242 | 0,4821 | 0,76445 |
| | Bagged Trees | 0,31653 | 0,22443 | 0,47374 | 0,77576 | 2800 | 0,30259 | 0,23895 | 0,48883 | 0,75783 |
| GPR | Squared Exponential GPR | 0,28555 | 0,14977 | 0,38701 | 0,85035 | 3700 | 0,24658 | 0,10322 | 0,32128 | 0,89539 |
| | Matern 5/2 GPR | 0,27717 | 0,14430 | 0,37986 | 0,85582 | 4000 | 0,24469 | 0,10427 | 0,32291 | 0,89432 |
| | Exponential GPR | 0,28547 | 0,15090 | 0,38845 | 0,84923 | 4500 | 0,25913 | 0,11794 | 0,34342 | 0,88047 |
| | Rational Quadratic GPR | 0,27595 | 0,143 | 0,37816 | 0,85712 | 4800 | 0,2453 | 0,10485 | 0,32381 | 0,89373 |
| NN | Narrow Neural Network | 0,32247 | 0,22607 | 0,47547 | 0,77411 | 5000 | 0,26544 | 0,11378 | 0,33732 | 0,88468 |
| | Medium Neural Network | 0,44 | 0,33093 | 0,57526 | 0,66934 | 5300 | 0,37638 | 0,25744 | 0,50738 | 0,73909 |
| | Wide Neural Network | 0,30638 | 0,16678 | 0,40839 | 0,83336 | 6600 | 0,31536 | 0,18774 | 0,43329 | 0,80973 |
| | Bilayered Neural Network | 0,33344 | 0,26627 | 0,51602 | 0,73395 | 6000 | 0,26429 | 0,13394 | 0,36598 | 0,86425 |
| | Trilayered Neural Network | 0,31347 | 0,22775 | 0,47723 | 0,77244 | 4100 | 0,256 | 0,15174 | 0,38953 | 0,84622 |
| KAR | SVM Kernel | 0,34911 | 0,2167 | 0,46551 | 0,78348 | 4100 | 0,30353 | 0,14815 | 0,38491 | 0,84985 |
| | Least Squares Regression Kernel | 0,38434 | 0,24147 | 0,49139 | 0,75873 | 3300 | 0,35455 | 0,18733 | 0,43282 | 0,81014 |

The RMSE and MAE metrics were used for comparison between the models. Figures 3 and 4 show the performance results obtained by each ML model during the training and test process. The changes and achievements are represented by different colors and symbols are shown on the graph. This graph provides a visual comparison of model performances by showing how successful each model is in the training process. Also, the Interactions Linear algorithm was removed from the graphs due to the presence of outliers.

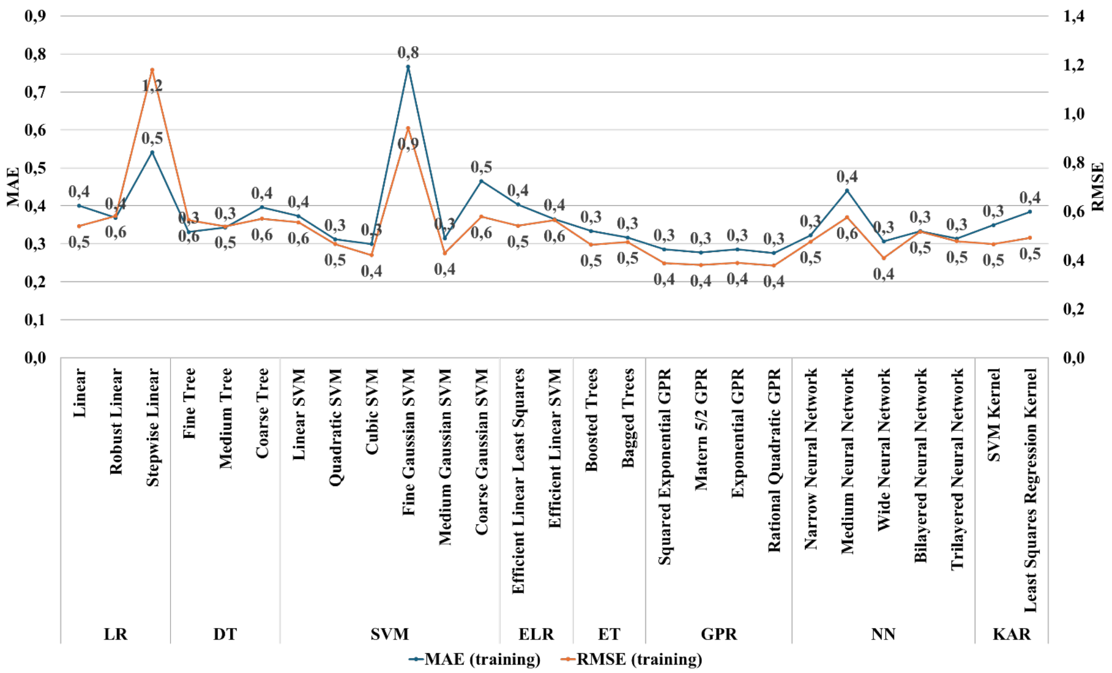


**Figure 3.** Training performance comparison for ML models

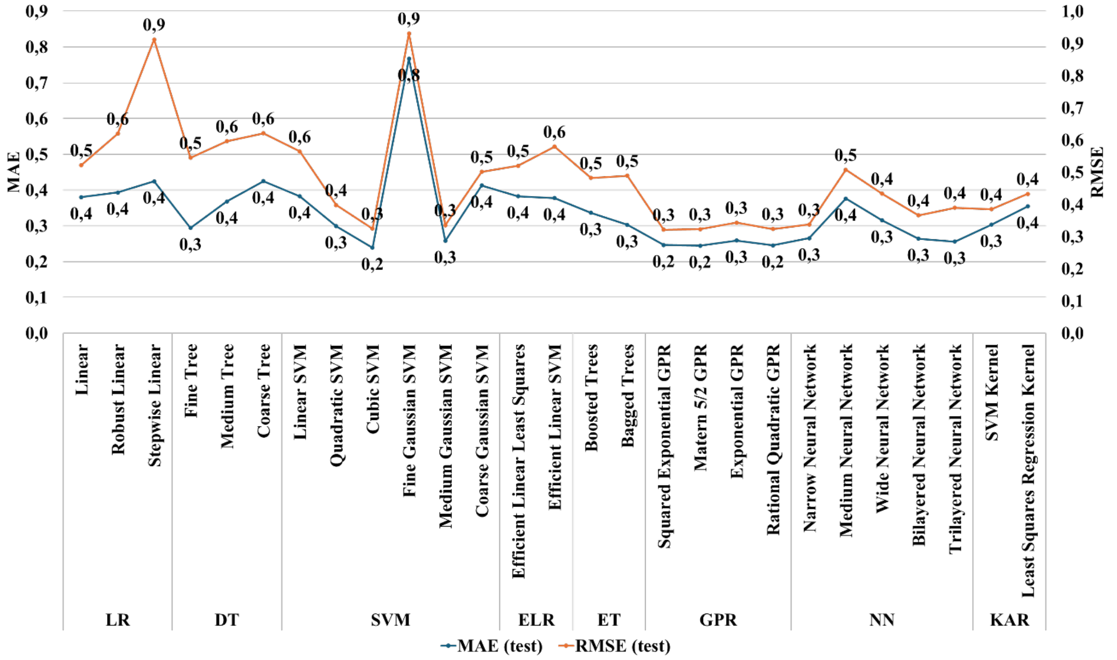


**Figure 4.** Test performance comparison for ML models

Figure 5 shows the training time of each ML model in observations per second (obs/sec). This figure provides a visual comparison of the variation and differences in the training time of each model. This graph visually highlights the presence of time performance between the models by representing the training times of each model with different columns. This distribution provides an important indication of the computational complexity of each algorithm and its adaptability to the dataset.

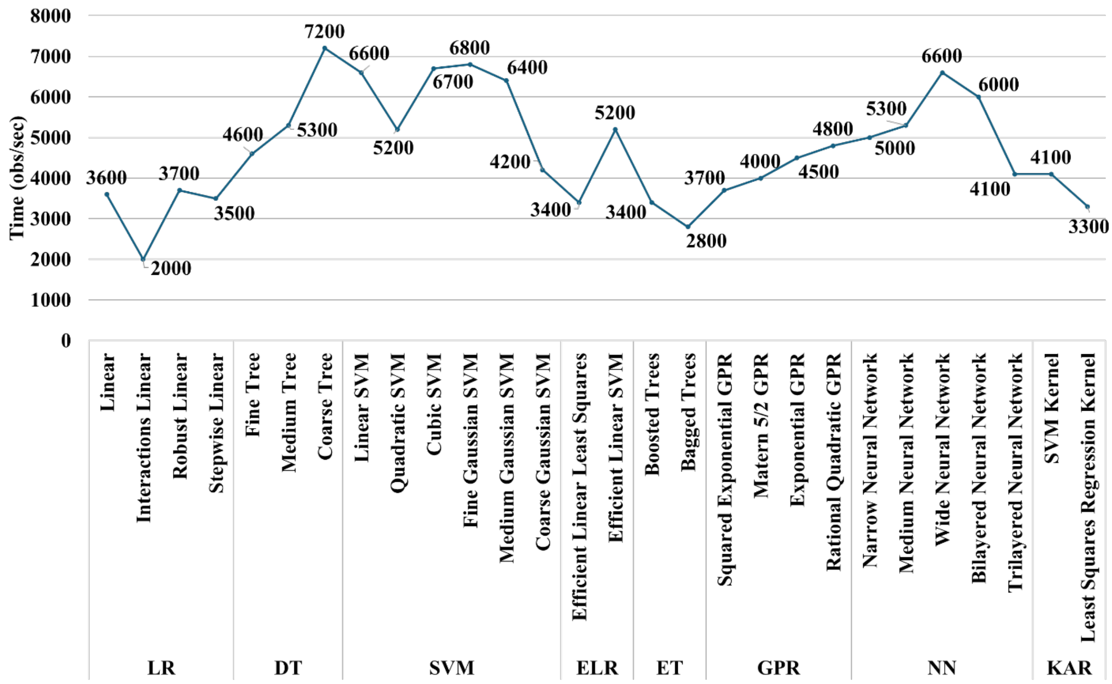


**Figure 5.** Training time distribution for all ML models

## 5. Discussion and Conclusion

This paper aims to predict ticket prices based on airline flight data using ML algorithms and to compare the performance of ML algorithms. The secondary objective of this paper is to identify the main factors affecting airline ticket prices. The flight and ticket price datasets of THY and PGS that were obtained from open-access sources are used in this paper. The final dataset consists of 962 records for 3 months from June 1st, 2022 to August 30th, 2022 and includes 19 different variables. In this paper, we compared various ML models to predict airline ticket prices, considering performance metrics such as MAE, MSE, RMSE, and $R^2$ during both training and test phases. According to the model training and test results, the best algorithm is GPR with $R^2$: 0.86 (training) and $R^2$: 0.90 (test). Our findings align with existing literature, reinforcing the efficacy of certain models over others in specific contexts. Below, the performance results concerning previous studies are discussed.

Janssen et al. (2014) used a linear quantile mixed regression model to predict the lowest ticket price before departure, reporting reasonable short-term performance but inefficient long-term results. Our study shows that the standard LR model achieved a reasonable $R^2$ value of 0.72 during testing, indicating good short-term prediction capabilities. However, the Interactions Linear model performed poorly with a negative $R^2$ value, suggesting its inefficiency in capturing complex price patterns over the long term.

Wohlfarth et al. (2011) recommended using CART and RF models for predicting the best time to buy tickets. Our results for the Medium Tree model show an $R^2$ of 0.71 during testing, supporting the effectiveness of decision tree approaches. Fine Tree and Coarse Tree models also performed adequately, though with slightly lower $R^2$ values. Also, Kalampokas et al. (2023) conducted a study to predict airfare prices using the decision tree model, achieving an average $R^2$ value of 0.83. Although our $R^2$ result is lower, it can be said that a result close and parallel to the literature has been obtained. The reason for this low result is mainly related to the quality and size of the dataset used. These findings collectively underscore the robustness and reliability of decision tree models in accurately forecasting airline ticket prices.

Tziridis et al. (2017) found that Bagging Regression Trees and RF models performed best among eight ML models, including SVMs. Similarly, Kalampokas et al. (2023) conducted a study to forecast airfare prices utilizing the RF model, achieving an average $R^2$ value of 0.87. In our analysis, the Cubic SVM model achieved an impressive $R^2$ of 0.89 during testing, outperforming other SVM variants. The Medium Gaussian SVM also showed strong performance with an $R^2$ of 0.88, confirming the potential of SVM models in airfare prediction. Also, Kalampokas et al. (2023) achieved a study to predict airline ticket prices using the SVM model, reaching an average $R^2$ value of 0.80. Furthermore, Gui et al. (2020) applied an ensemble model combining random tree models with DL for flight delay prediction, achieving high accuracy. Similarly, our study indicates that Boosted Trees and Bagged Trees performed well, with the Boosted Trees model achieving an $R^2$ of 0.78. These results corroborate the utility of ensemble methods in handling complex, non-linear relationships in airfare data.

DL models, such as those proposed by Tziridis et al. (2017) and various other studies, have shown significant promise. Likewise, Kalampokas et al. (2023) conducted a study to forecast airfare prices utilizing the Multi-Layer Perceptron (MLP) model, achieving an average $R^2$ value of 0.91. Our Narrow Neural Network model achieved an $R^2$ of 0.88, and the Bilayered Neural Network performed well with an $R^2$ of 0.86. These findings are consistent with the literature, indicating that neural networks are highly effective in capturing intricate patterns in ticket pricing data. Our study's GPR models, particularly the Squared Exponential GPR and Matern 5/2 GPR, demonstrated high performance with $R^2$ values of 0.895 and 0.894, respectively. These results align with existing research suggesting that GPR models are well-suited for regression tasks involving complex, non-linear data.

Across different model categories, this paper confirms that advanced ML techniques, especially ensemble models and neural networks, offer superior performance in predicting airline ticket prices. Models such as the Cubic SVM and GPR variants showed outstanding predictive accuracy, with $R^2$ values close to or exceeding 0.89. This indicates their robustness and adaptability to the dynamic nature of airfare pricing.

On the other hand, it is possible to note several key factors influencing ticket prices. There is a positive correlation between the distance and duration of the flight and ticket prices; for instance, flights over 300 minutes exhibit an average price increase of 15% compared to shorter duration. Additionally, longer flight durations increase ticket prices due to increased operational costs. Also, seasonal demand can significantly impact ticket prices. Furthermore, three monthly demand fluctuations reveal that prices increase by the day. The presence of competing airlines on the same route reduces ticket prices and our observations indicate that routes served by three or more airlines tend to have lower prices due to increased competition. Besides, fluctuations in fuel prices can directly impact ticket pricing. Macroeconomic factors such as Gross Domestic Product growth and inflation rates also affect ticket prices, as higher disposable incomes during periods of economic growth lead to increased demand for air travel, subsequently driving up prices, while during economic downturns, prices tend to stabilize or decrease. By considering these factors, airlines can better understand and strategically manage their ticket pricing to optimize revenue and market competitiveness.

Based on our analysis, several practical recommendations can be made for airlines to optimize their pricing strategies. Firstly, airlines should adopt dynamic pricing models that adjust ticket prices based on real-time demand and supply conditions. Implementing sophisticated algorithms that consider factors such as booking time, competition, and customer behavior can help maximize revenue. Secondly, strategic route planning is crucial; airlines should strategically plan routes to capitalize on high-demand periods and destinations by offering more flights during peak seasons and reducing frequency during off-peak times, thus optimizing operational efficiency and profitability. To mitigate the impact of volatile fuel prices, airlines can engage in fuel hedging, a financial strategy that involves locking in fuel prices for future purchases, providing cost predictability, and reducing the risk of sudden price hikes. Additionally, developing robust loyalty programs can help airlines retain customers and encourage repeat business; offering incentives such as discounts, upgrades, and exclusive benefits can enhance customer satisfaction and brand loyalty. Lastly, leveraging advanced technologies such as AI and big data analytics can improve operational efficiency; predictive maintenance, optimized flight paths, and automated customer service are areas where technology can significantly reduce costs and improve the quality of service. By implementing these recommendations, airlines can strategically manage their pricing, enhance customer loyalty, and achieve greater operational efficiency.

In conclusion, our findings support the existing literature, demonstrating that ML models, particularly ensemble methods and DL models, significantly enhance the accuracy of airline ticket price predictions. These models provide valuable insights and practical tools for airlines to optimize their pricing strategies, ultimately contributing to more efficient revenue management in the highly competitive aviation industry. Future research should continue exploring these advanced techniques, incorporating larger and more diverse datasets to further refine and validate model performance.

## 6. Limitations and Future Research

Our analysis has certain limitations and suggests several directions for future research. Firstly, data limitations exist as our analysis was based on data from a limited number of airlines and routes, which may not fully represent the global airline industry. Future studies should aim to include a more comprehensive dataset. Additionally, unforeseen external factors such as geopolitical events, pandemics, and natural disasters can significantly impact ticket prices, and these factors were not fully accounted for in our analysis, and thus should be considered in future research. Moreover, rapid technological advancements in aircraft efficiency and alternative fuels were not deeply explored in this study, yet these factors could have significant implications for future ticket pricing strategies.

Looking ahead, future research should aim to conduct a more comprehensive global analysis, including a broader range of airlines and routes, to provide a more holistic understanding of the factors affecting ticket prices. Additionally, investigating the impact of emerging technologies such as electric aircraft, autonomous flights, and sustainable fuels on ticket pricing can provide valuable insights for the industry. Exploring consumer behavior patterns, including preferences and purchasing behaviors, can help airlines develop more targeted and effective pricing strategies. Furthermore, conducting longitudinal studies to track changes in ticket pricing over time and across different economic cycles can provide deeper insights into the dynamic nature of airline pricing. By addressing these limitations and pursuing these research directions, future studies can offer a more robust and detailed understanding of airline ticket pricing dynamics.

**Peer Review:** Externally peer-reviewed.
**Conflict of Interest:** Author declared no conflict of interest.
**Financial Disclosure:** Author declared no financial support.

**ORCID IDs of the authors**

Hüseyin Korkmaz 0000-0002-2438-6919

## REFERENCES

Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F. (2021). Airline ticket price and demand prediction: A survey. *Journal of King Saud University - Computer and Information Sciences, 33*(4), 375–391. https://doi.org/10.1016/j.jksuci.2019.02.001

Aliberti, A., Xin, Y., Viticchié, A., Macii, E., & Patti, E. (2023). Comparative analysis of neural networks techniques to forecast Airfare Prices. *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 1023–1029.

https://doi.org/10.1109/COMPSAC57700.2023.00157

Barrett, G. B. (2000). The Coefficient of Determination: Understanding r squared and R squared. *The Mathematics Teacher, 93*(3), 230–234.

Chen, T., Yin, H., Chen, H., Wu, L., Wang, H., Zhou, X., & Li, X. (2018). TADA: Trend Alignment with Dual-Attention Multi-task Recurrent Neural Networks for Sales Prediction. *2018 IEEE International Conference on Data Mining (ICDM)*, 49–58. https://doi.org/10.1109/ICDM.2018.00020

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Deng, T. (2024). International flight fare prediction and analysis of factors impacting flight fare. *Theoretical and Natural Science, 31*(1), 329–335. https://doi.org/10.54254/2753-8818/31/20241079

Di Bucchianico, A. (2008). Coefficient of determination (R 2). *Encyclopedia of Statistics in Quality and Reliability*.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics, 49*(1), 92. https://doi.org/10.2307/1937887

Gibbons, J. D. (1997). Nonparametric methods for quantitative analysis. *Journal of the Operational Research Society, 48*(8), 853.

Gordiievych, A., & Shubin, I. (2015). Forecasting of airfare prices using time series. *2015 Information Technologies in Innovation Business Conference (ITIB)*, 68–71. https://doi.org/10.1109/ITIB.2015.7355055

Groves, W., & Gini, M. (2015). On Optimizing Airline Ticket Purchase Timing. *ACM Transactions on Intelligent Systems and Technology, 7*(1), 1–28. https://doi.org/10.1145/2733384

Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2020). Flight Delay Prediction Based on Aviation Big Data and Machine Learning. *IEEE Transactions on Vehicular Technology, 69*(1), 140–150. https://doi.org/10.1109/TVT.2019.2954094

Howell, D. C. (1992). *Statistical methods for psychology*. PWS-Kent Publishing Co.

Huang, H.-C. (2013). A Hybrid Neural Network Prediction Model of Air Ticket Sales. *TELKOMNIKA Indonesian Journal of Electrical Engineering, 11*(11). https://doi.org/10.11591/telkomnika.v11i11.2762

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Janssen, T., Dijkstra, T., Abbas, S., & van Riel, A. C. (2014). A linear quantile mixed regression model for prediction of airline ticket prices. *Radboud University, 3*.

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

Karataş, G. (2021). The Effects of Normalization and Standardization an Internet of Things Attack Detection. *European Journal of Science and Technology, 29*, 187–192. https://doi.org/10.31590/ejosat.1017427

Kumar, A. (2023). *Airline Price Prediction Using XGBoost Hyper-parameter Tuning* (pp. 239–248). https://doi.org/10.1007/978-3-031-28183-9_17

Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling Long- and Short-Term Temporal Patterns with Deep Neural Networks. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104. https://doi.org/10.1145/3209978.3210006

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting, 1*(2), 111–153. https://doi.org/10.1002/for.3980010202

Miles, J., & Banyard, P. (2007). *Understanding and using statistics in psychology: A practical introduction*.

Nevitt, J., & Hancock, G. R. (2000). Improving the Root Mean Square Error of Approximation for Nonnormal Conditions in Structural Equation Modeling. *The Journal of Experimental Education, 68*(3), 251–268. https://doi.org/10.1080/00220970009600095

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. W. (2017). A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2627–2633. https://doi.org/10.24963/ijcai.2017/366

Ren, R., Yang, Y., & Yuan, S. (2014). Prediction of airline ticket price. *University of Stanford*, 1–5.

Sammut, C., & Webb, G. I. (2010). Mean absolute error. *Encyclopedia of Machine Learning, 652*.

Santana, E., Mastelini, S., & Jr., S. (2017). Deep Regressor Stacking for Air Ticket Prices Prediction. *Anais Do Simpósio Brasileiro de Sistemas de Informação (SBSI)*, 25–31. https://doi.org/10.5753/sbsi.2017.6022

Sherly Puspha Annabel, L., Ramanan, G., Prakash, R., & Sreenidhi, S. (2023). *Machine Learning-Based Approach for Airfare Forecasting* (pp. 901–912). https://doi.org/10.1007/978-981-19-6634-7_65

Shih, S.-Y., Sun, F.-K., & Lee, H. (2019). Temporal pattern attention for multivariate time series forecasting. *Machine Learning, 108*(8–9), 1421–1441. https://doi.org/10.1007/s10994-019-05815-0

Tziridis, K., Kalampokas, Th., Papakostas, G. A., & Diamantaras, K. I. (2017). Airfare prices prediction using machine learning techniques. *2017 25th European Signal Processing Conference (EUSIPCO)*, 1036–1039. https://doi.org/10.23919/EUSIPCO.2017.8081365

Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S.-C. (2019). A Framework for Airfare Price Prediction: A Machine Learning Approach. *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, 200–207. https://doi.org/10.1109/IRI.2019.00041

Wohlfarth, T., Clemencon, S., Roueff, F., & Casellato, X. (2011). A Data-Mining Approach to Travel Price Forecasting. *2011 10th International Conference on Machine Learning and Applications and Workshops*, 84–89. https://doi.org/10.1109/ICMLA.2011.11

Yujing, D., Zhihao, W., & Youfang, L. (2020). Flight passenger load factors prediction based on RNN using multi-granularity time attention. *Computer Engineering, 46*(509), 01.

Zhao, Z., You, J., Gan, G., Li, X., & Ding, J. (2022). Civil airline fare prediction with a multi-attribute dual-stage attention mechanism. *Applied Intelligence, 52*(5), 5047–5062. https://doi.org/10.1007/s10489-021-02602-0

### How cite this article

Korkmaz, H. (2024). Prediction of airline ticket price using machine learning method. *Journal of Transportation and Logistics 9*(2), 205-218. https://doi.org/10.26650/JTL.2024.1486696