International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

Unveiling the Dynamics of Airfare - A Machine LearningApproach to Price Prediction

[1] A. HARIDWATHI, [2] M. NISHITH CHOWDARY, [3] D. HARSHITA, [4] A. CHARAN TEJA,
[5] Dr. K. VASANTH KUMAR, [6] Mr. LALAM RAMU

Department of Computer Science and Engineering – Internet of ThingsMalla Reddy Engineering College, Hyderabad,
Telangana nishith374@gmail.com

Abstract**—** This paper focuses on predicting airfare prices by identifying key factors that influence ticket costs for flights. Eight advanced machine learning models are used to forecast these prices based on the selected features. The paper compares the performance of these models in terms of prediction accuracy and investigates how this accuracy varies with different sets of flight details. The experiments utilize a unique dataset comprising 1,814 flights operated by Aegean Airlines on a specific internationalroute (from Thessaloniki to Stuttgart) to train each machine learning model. The findings demonstrate that these models can effectively address the regression problem, achieving an accuracy level of nearly 88% when certain flight details are considered.

Keywords—machine learning models; prediction model; airfare price; pricing models.

I. INTRODUCTION

Nowadays, airlines use various techniques and procedures to allocate airfare pricing in a dynamic manner [1], [2]. These tactics consider a variety of financial, marketing, commercial, and social issues that are directly related to ultimate flight pricing.

Because of the tremendous complexity of the pricing algorithms used by airlines, it is extremely difficult for a clientto acquire an air ticket at the lowest possible price, as prices fluctuate frequently.

As a result, numerous approaches [3], [4] that may anticipate the airfare price and offer the consumer with the optimal time to acquire an air ticket have recently been presented. The bulk of these technologies make use of sophisticated prediction models developed in the computational intelligence study field known as Machine Learning.

More specifically, Groves and Gini [4] used a PLS regression model to optimize airline ticket purchases with 75.3% accuracy. Papadakis [5] predicted whether the ticket price will fall in the future by treating the situation as a classification job with Ripple Down Rule Learner (74.5% accuracy), Logistic Regression (69.9% accuracy), and Linear SVM (69.4% accuracy). Janssen [6] suggested a linear quantile mixed regression model to forecast air ticket prices with satisfactory performance for inexpensive tickets over several days before departing. Ren, Yang, and Yuan [7] compared Linear Regression (77.06% accuracy), NaDve Bayes (73.06% accuracy), Softmax Regression (76.84% accuracy), and SVM (80.6% accuracy for two bins) models for forecasting airline ticket prices.

All of the aforementioned research used just a small number of ML models, primarily classical ones, to forecast airline travel costs throughout the world. However, to the best of the authors' knowledge, the performance of cutting-edge machine learning models on this subject has yet to be investigated.The proposed paper's contribution is stated as follows: (1) the first forecast of flight pricing in Greece, (2) an examination of the factors that impact airfare prices, and (3) a performance analysis of cutting-edge ML models in airfare prediction.While classical ML models have been commonly used in previous studies, the utilization of state-of-the-art ML techniques like deep learning models, ensemble methods, and advanced feature engineering approaches has not been extensively explored in this domain.

By evaluating these cutting-edge models, the paper aims to provide insights into their efficacy, scalability, and potential advantages over traditional approaches in accurately forecasting airfare prices. This involves not only considering traditional features like flight duration, time of booking, and seat class but also exploring more nuanced factors such as seasonal demand fluctuations, economic indicators, and even geopolitical events that may influence travel patterns and pricing dynamics.

The rest of the paper is arranged as follows: Section II providessome basic information on machine learning and how it might beapplied to the problem of predicting airline prices. Section III provides a theoretical overview of the present study, while Section IV details the experimental technique and outcomes of the models utilized. Finally, Section V summarizes the whole study and suggests some research possibilities for future work.

II. MACHINE LEARNING

One of the hottest study areas in computer science and engineering right now is machine learning, which has applications across many academic fields. It offers a selection oftechniques, tools, and algorithms that enable machines to exhibitintelligence.

The modeling tools that machine learning (ML) offers are powerful because they can be taught with a collection of data that describes a particular problem through a learning process and can then respond to comparable, unseen data in a common way.

As the volume of data used for learning increases, some well- known machine learning models include Multilayer Perceptrons (MLPs), Radial Basis Function (RBF) and Generalized Regression (GRNN) neural networks [8], Support Vector Machines (SVMs) [8], Decision Trees (DTs) [9], Extreme Learning Machines (ELMs) [10], etc. However, the increase of the training data needs parallel implementations [11] of the ML algorithms using specialized software and hardware platforms.

There are two ways to approach the problem of predicting flight prices in the context of machine learning. The firstmethod approaches the issue of predicting the price of airline tickets as a regression problem, whereas the second method approaches it as a classification challenge. Since regression models attempt to estimate a function that specifies the mapping rule between data attributes and airfare prices, the former technique is often used for the precise prediction of the price of an airline ticket. Although the latter method cannot pinpoint an exact airfare, it can offer guidance on price ranges and whether to purchase a ticket at a certain price or not.

Since there hasn't been much focus on assessing the most advanced regression machine learning models for that issue, this study considers the first example of flight price prediction via regression.

III. Current Study

The Greek airline Aegean Airlines [12] and its trip from Thessaloniki to Stuttgart are first chosen as the subject of our inquiry's case study.

The current study is divided into four main stages: (1) choosing the flight characteristics that affect airfare prices; (2) gathering sufficient flight data to train and test the applied machine learning models; (3) choosing the regression ML models to compare; and (4) conducting an experimental evaluation of the ML models.

Each processing phase is discussed in more detail in the following:

Phase 1 (Feature Selection) - During this phase the most informative features of a flight that determine the prices of the air tickets are decided. This phase is very important since it defines the problem under solving.

For every flight the following features were considered:

- F1: Feature 1 - Airline.
- F2: Feature 2 – Date of Journey.
- F3: Feature 3 - Source.
- F4: Feature 4 - Destination.
- F5: Feature 5 - Route.
- F6: Feature 6 – Departure Time.
- F7: Feature 7 – Arrival Time.
- F8: Feature 8 - Duration.
- F9: Feature 9 – Total Stops
- F10: Feature 10 – Additional Info
- F11: Feature 11- Price

The study's robustness is increased by the one-leave-out procedure, which systematically removes significant information one at a time while assessing each feature's impact on prediction accuracy. The days that pass between the time of purchase and the flight—feature F2—are crucial because they illustrate how dynamic pricing works. In order to help airlines and consumers make judgments regarding pricing and scheduling, this rule's research of F2's effect can shed information on how booking time influences airfareforecasts.

Phase 2 (Data Collection) - In this study, our interest is focused on the prediction of a single airfare price without return. For the sake of the experiments a set of flights to the same destination (from Thessaloniki to Stuttgart) for the period between December and July, is collected. For each flight the eleven features (F1:F11) were manually collected from the Web, 1814 flights were recorded totally and are available in [13].

Phase 3 (ML Models Selection) - Eight state of the art regression ML models [8], [10], [14], [15], [16] were selected for the current study and applied to the same data of flights. The ML models compared in this work are the following:

- ○ Multilayer Perceptron (MLP).
- ○ Generalized Regression Neural Network.
- ○ Extreme Learning Machine (ELM).
- ○ Random Forest Regression Tree.
- ○ Regression Tree.
- ○ Bagging Regression Tree.
- ○ Regression SVM (Polynomial and Linear).
- ○ Linear Regression (LR).

Phase 4 (Evaluation) - The ML models stated above weretrained using a 10-fold cross-validation process using the 1814 flights that were gathered during phase 2. The prediction accuracy (% - MSE between the intended and predicted prices) and training time (in seconds) are the performance metrics that are used to compare the models.

IV. Simulations

For the sake of the experiments a set of simulations were arranged and executed under the MATLAB environment in a i5-750 2.67 GHz PC with 8GB memory. The configuration of the ML models was decided by applying grid search and is summarized in Table I.

TABLE I. Models Configuration

ML Model	***Configuration***
Multilayer Perceptron (MLP)	3 hidden layers 5 nodes each layer
Generalized Regression Neural Network	spread=1.0
Extreme Learning Machine	10 neurons
Random Forest Regression Tree	300 weak classifiers (decision trees)

Regression Tree	MinParentSize=10
	MinLeafSize=3
	MaxNumSplits=45

Bagging Regression Tree	500 weak classifiers (decision trees)
Regression SVM (Polynomial)	order=3
Regression SVM (Linear)	stochastic gradient descent solver
Linear Regression	dual stochastic gradient descent solver

Every experiment was subjected to a 10-fold cross-validation process, and the average performance of every model is shownin this section.Table II displays the performance of all models for the case of the full feature set (eleven features), with bold faced being the best-performing model.

TABLE II. RESULTS WITH ALL FEATURES

ML Model	***Accuracy (%)***	***Execution Time (sec)***
Multilayer Perceptron	80.28	20.88
Generalized Regression Neural Network	66.83	0.13
Extreme Learning Machine	68.68	0.05
Random Forest Regression Tree	85.91	5.50
Regression Tree	84.13	0.04
Bagging Regression Tree	**87.42**	17.05
Regression SVM (Polynomial)	77.00	1.23
Regression SVM (Linear)	49.40	0.34
Linear Regression	57.25	0.10

From the results of Table II it is obvious that the “Bagging Regression Tree” model outperforms the other models, while its training is quite fast. Moreover, the “Random Forest Regression Tree” seems to be an alternative choice since it shows similar performance in less time.

In order to analyze the influence of the used features to the prediction accuracy of the models, the same experiment is repeated several times by leaving out some features, one at a time. In this context the first two time features were removed and the experiment is repeated with six features (F3:F8). The corresponding results are presented in Table III.

TABLE III. RESULTS WITHOUT F1 & F2 FEATURES

ML Model	***Accuracy (%)***	***Execution Time (sec)***
Multilayer Perceptron	75.50	19.90
Generalized Regression Neural Network	65.72	0.11
Extreme Learning Machine	67.58	0.04
Random Forest Regression Tree	**79.40**	10.6
Regression Tree	78.76	0.06
Bagging Regression Tree	77.50	15.07
Regression SVM (Polynomial)	76.18	1.10
Regression SVM (Linear)	48.50	0.35
Linear Regression	56.20	0.23

From the above results it is obvious that almost all models shown lower (up to 10%) prediction accuracy and greater execution time. These results reveal that the timing features “departure time” and “arrival time” influence significantly the airfare prices. Furthermore, the increase of the execution time means that the training procedure converges quite later for almost all the models.

Table IV summarizes the performance of the models when the “duration” feature (F8) is omitted during training.

TABLE IV. RESULTS WITHOUT F8 FEATURE

ML Model	***Accuracy (%)***	***Execution Time (sec)***
Multilayer Perceptron	81.58	5.65
Generalized Regression Neural Network	66.83	0.32
Extreme Learning Machine	66.88	0.086
Random Forest Regression Tree	86.18	5.28
Regression Tree	84.22	0.02
Bagging Regression Tree	**87.59**	13.73
Regression SVM (Polynomial)	79.38	0.98
Regression SVM (Linear)	60.64	0.02
Linear Regression	57.07	0.05

In this case, we observe that all models were not affected as much as previously, except “Regression SVM” with Linear kernel. Therefore, one can conclude that the “day of week” does not influence airfare prices.

Table V, presents the performance of the models without using the “arrival time” feature (F7). The outcomes of this experiment reveal that this feature is not related with the price of the air ticket, since the models perform similarly or even worse with the case of using all features. Only the “Multilayer Perceptron” and the “Regression SVM” with Linear Kernel seems to be affected significantly by this feature.

TABLE V. RESULTS WITHOUT F7 FEATURE

ML Model	***Accuracy (%)***	***Execution Time (sec)***
Multilayer Perceptron	72.8	5.98
Generalized Regression Neural Network	66.14	0.34
Extreme Learning Machine	64.88	0.06
Random Forest Regression Tree	86.15	6.15
Regression Tree	84.22	0.059
Bagging Regression Tree	**87.93**	15.34
Regression SVM (Polynomial)	77.91	0.17
Regression SVM (Linear)	57.69	0.06
Linear Regression	57.92	0.02

Next, we are leaving out the “departure time” feature (F6), and the models are executed again. Their performance is similar with that of the first experiment, as illustrated in Table VI.

TABLE VI. RESULTS WITHOUT F6 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	77.94	5.74
Generalized Regression Neural Network	66.31	0.25
Extreme Learning Machine	68.5	0.05
Random Forest Regression Tree	86.17	5.54
Regression Tree	84.13	0.02
Bagging Regression Tree	**87.60**	16.47
Regression SVM (Polynomial)	67.2	0.15
Regression SVM (Linear)	57.69	0.05
Linear Regression	57.92	0.02

The "Bagging Regression Tree" outperforms all the models not only in this experiment, but also all the models under different feature sets examined previously. The reminder models seem not to be affected by the exclusion of "holiday day" feature.

The last experiment is executed without using the "route" feature (F5), with similar results with the first experiment.

TABLE VII. RESULTS WITHOUT F5 FEATURE

ML Model	*Accuracy (%)*	*Execution Time (sec)*
Multilayer Perceptron	78.62	7.43
Generalized Regression Neural Network	65.24	0.32
Extreme Learning Machine	66.83	0.03
Random Forest Regression Tree	86.04	4.79
Regression Tree	83.88	0.01
Bagging Regression Tree	**87.91**	16.32
Regression SVM (Polynomial)	77	0.14
Regression SVM (Linear)	49.4	0.05
Linear Regression	57.25	0.02

Concluding the previous study, one can claim that "Bagging Regression Tree", "Random Forest Regression Tree", "Regression Tree" and MLP models are the most stable models according to their accuracy scores. In addition, as far as the execution time is concerned the best models are "Random Forest Regression Tree" and "Regression tree".

V. CONCLUSION

A preliminary investigation on "airfare prices prediction" was included in this article. We collected airfare data from the internet for a particular Greek airline company, Aegean Airlines, and demonstrated that it is possible to forecast trip costs using past pricing data. The findings of the experiment demonstrate that ML models are a useful tool for estimating the cost of flights. The data gathering and feature selection, from which we extracted some insightful insights, are additional crucial elements in airfare prediction. We deduced from the tests which features had the most impact on airfare prediction.

Other factors exist that might increase the prediction accuracy in addition to the ones that were chosen. This study might be expanded in the future to forecast the airfare costs for the airline's whole flight schedule. Although more research on bigger airfare data sets is necessary, this first pilot study demonstrates how machine learning models might help customers buy tickets at the optimal time of year.

REFERENCES

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

[2] R. Redondi, P. Malighetti, and S. Paleari, "Has Ryanair's pricing approach altered over time? An empirical study of its flights from 2006to 2007 was published in Tourism Management, vol. 31, no. 1, 2010, pp.36–44.

[3] Groves, W. and Gini, M., "A regression model for predicting optimal timing for airline ticket purchases," University of Minnesota, Minneapolis, Technical Report 11-025, 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 06 - 10, 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 Univerisy, 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] A. Lakshmanarao, Srisaila A2, Srinivasa Ravi Kiran T, Lalitha G,

Vasanth Kumar K., “Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms”, iJOE – Vol. 18, No. 13, 2022, https://doi.org/10.3991/ijoe.v18i13.33315.

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

[13] 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.

[16] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola and V. Vapnik, "Support vector regression machines," Advances in neural information processing systems, vol. 9, pp. 155-161, 1997.