International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 01 | Jan - 2026 SJIF Rating: 8.586 ISSN: 2582-3930

Intelligent Flight Ticket Price Prediction Using Data-Driven Machine Learning Models

Souvik Dey[1], Sudipta Kumar Dutta[2]

[1]*Souvik Dey, B. Tech in Computer Science Engineering from B.P. Poddar Institute of Management and Technology*

[2]*Mr. Sudipta Kumar Dutta, Department of CSE at B.P. Poddar Institute of Management and Technology*

Abstract - This paper deals with the problem of airfare prices prediction. For this purpose, a set of features characterizing a typical flight is decided, supposing that these features affect the price of an air ticket. The features are applied to eight states of the art machine learning (ML) models, used to predict the air tickets prices, and the performance of the models is compared to each other. Along with the prediction accuracy of each model, this paper studies the dependency of the accuracy on the feature set used to represent an airfare. For the experiments a novel dataset consisting of 1814 data flights of the Aegean Airlines for a specific international destination (from Thessaloniki to Stuttgart) is constructed and used to train each ML model. The derived experimental results reveal that the ML models are able to handle this regression problem with almost 88% accuracy, for a certain type of flight features.

Key Words: optics, photonics, light, lasers, templates, journals

❑ LITERATURE SURVEY

Flight fare prediction has emerged as an important research area due to the dynamic, non-linear behavior of airline pricing systems. Airfare varies with factors such as time of booking, seasonality, travel route, airline policies, and market demand. The reviewed literature from multiple studies highlights various machine learning approaches used to model this complex pricing structure.

1. Early Modelling of Airline Price Behavior

Initial studies examined **general price behavior**, showing that airlines change fares depending on weekdays, seasons, and proximity to the departure date. These works analysed historical fare patterns and showed that models must account for high volatility, non-linearity, and heterogeneous pricing schemes.

Some works employed **non-parametric regression** to study optimal purchase timing. These studies revealed that prices often rise sharply as the journey date approaches, although exceptions exist depending on route and carrier. Such findings established the need for predictive systems capable of handling varied time-dependent patterns.

2. Statistical and Regression-Based Prediction Approaches

Several studies focused on **linear models**, such as:

- **Linear Regression**
- **PLSR (Partial Least Square Regression)**
- **Quantile-based regression**

These models worked reasonably well when relationships between variables were linear. However, they struggled to capture the deeper complexities of airfare behavior, particularly close to the date of departure. These studies also emphasized key predictive variables such as:

- Days left before departure
- Day of the week
- Seasonal patterns
- Historical price trends

Regression-based work laid the foundation for understanding which features are most influential but highlighted the limitations of simple statistical models for real-world price forecasting.

3. Machine Learning Models for Fare Prediction

More recent research transitioned toward **supervised machine learning**. Several studies benchmarked algorithms such as:

Decision Trees

Used due to interpretability and ability to handle mixed data types. Trees capture non-linear relationships and perform well for small datasets. However, they tend to overfit and yield lower accuracy than ensemble methods.

Random Forest Regressor

Many studies concluded that Random Forests outperform standalone decision trees due to ensemble averaging, robustness, and ability to handle high-dimensional data. Random Forests consistently achieved strong performance metrics such as:

- High R^2 scores
- Lower MSE and RMSE

- Better generalization on unseen data

Support Vector Machines

Applied in some studies, especially in earlier literature. SVM performed well for limited data but required careful tuning and scaling.

K-Nearest Neighbors (KNN)

KNN proved effective for capturing local patterns in price fluctuations, particularly when routes or seasons exhibited distinct clusters of behavior.

Artificial Neural Networks (ANN)

ANN-based models demonstrated the strongest performance among learning methods in some studies, especially when large datasets were used. Neural networks successfully captured complex relationships using multi-layer architectures and non-linear activation functions.

4. Data Handling and Feature Engineering

Across the reviewed studies, data preprocessing was highlighted as a crucial step:

- Removal of missing and duplicate entries
- Outlier detection (often using IQR methods)
- Transformation of categorical features via **one-hot encoding**
- Conversion of date and time variables into useful features
- Examination of correlations between price and factors such as duration, number of stops, and airline

Studies frequently used datasets sourced from Kaggle, scraping services, or airline APIs. Typical dataset features included:

- Airline name
- Source and destination
- Date of journey
- Departure/arrival times
- Duration and number of stops
- Route
- Additional info

These features formed the basis for most prediction models.

5. Comparative Evaluations

Multiple studies provided detailed comparisons of commonly used algorithms. The consensus is that:

- **Random Forest performs better** than traditional decision trees due to reduced overfitting.
- **ANN provides the strongest overall performance**, especially with large datasets.
- **Linear regression serves as a baseline model**, but is insufficient alone for accurate prediction.
- **KNN adds value in recognizing route-specific or seasonal clusters.**

Performance metrics typically included:

- **R^2 score**
- **MAE**
- **MSE / RMSE**
- **MAPE**

Across studies, Random Forest and ANN consistently achieved lower prediction errors and higher explanatory power.

6. Visualization and Exploratory Findings

Before modelling, many studies conducted exploratory data analysis revealing patterns such as:

- Prices change significantly with flight duration.
- More stops generally correlate with different fare tiers.
- Certain airlines show consistently higher average prices.
- Departures during peak times often exhibit higher fares.
- Outliers in price data must be treated to avoid model distortion.

Visualization techniques such as box plots, scatter plots, mutual information charts, and distribution graphs were widely used.

7. Recent Developments and Integrated Systems

Recent studies have integrated machine learning models into:

- Flight price recommendation engines
- Optimal purchase timing systems
- Real-time airfare monitoring dashboards

These works combine decision trees, Random Forests, KNN, and linear regression to build more robust and adaptable fare prediction solutions. The focus is shifting toward **hybrid models**, improved feature selection, and deployment of prediction systems in real-world applications.

1. INTRODUCTION

Airfare prices are known to fluctuate dynamically, and passengers who have booked flight tickets in the past are well aware of these rapid changes. Airlines employ advanced Revenue Management strategies to implement dynamic pricing models that adjust fares based

on various factors **[2]**. The lowest available ticket price at any given moment may vary significantly over time, as airfare can increase or decrease depending on demand patterns, market behavior, and prediction models discussed in prior research **[1], [3], [4], [7]**. These pricing systems automatically modify fares according to the time of day—such as morning, afternoon, or night—and may also fluctuate across seasons, including winter, summer, and festive periods **[2], [3]**. While the primary objective of airlines is to maximize revenue, passengers consistently seek the lowest possible fare **[4], [6]**.

Many travelers believe that purchasing tickets well in advance of the departure date guarantees a lower price. However, studies show that this assumption is not always accurate, and passengers may sometimes end up paying more than necessary for the same seat **[1], [5], [7], [9]**. Reports further indicate that India's civil aviation sector is experiencing rapid growth. India ranked as the third-largest aviation market in 2020 and is projected to become the largest by 2030 **[8]**. Air traffic in India was expected to exceed 100 million passengers by 2017, up from 81 million in 2015. Moreover, Google search trends show that the phrase "Cheap Air Tickets" remains one of the most frequently searched terms related to air travel in India, reflecting consumers' ongoing efforts to find affordable fares **[2], [3], [8], [9]**. Overall, the percentage of flight tickets purchased at the lowest available price continues to increase as travelers become more price-sensitive and informed **[5], [6], [7]**.

2. DATA-SET

The most appropriate dataset which is available to the public, was found in an online databank named Kaggle. The dataset is made up of information from over 300,000 past flight bookings from flights between 6 different cities in India. It consists of 9 independent variables, which are both numerical and categorical, and the dependent variable, being the ticket price in Indian Rupees. The independent variables include some of the previously named significant factors, such as the days left until the flight, the length of the flight, the number of layovers, and the travel class. These all provide a good base to train an ML model, as they should correlate well with the final ticket price and allow for accurate comparison between different models. However, a limitation of the dataset is that it does not include the number of seats left, giving no insights into the levels of supply and demand of the flights. Oil prices are also not included in the dataset, however, they may not be as significant in this investigation, as all flights are operated in India and range over a short time, making it safe to assume that prices maintained a steady level for all flights. This means that although not all factors are considered, the dataset includes a majority of price-defining factors, providing an adequate base to train ML models on. The figure-1 is Sample Data Set.

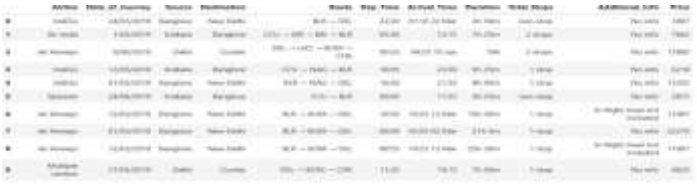


Figure -2: It is scattering plot and based on scatter plot. we can see changes of Ticket prices on Airline, Source, Destination, Total Stops, additional info.

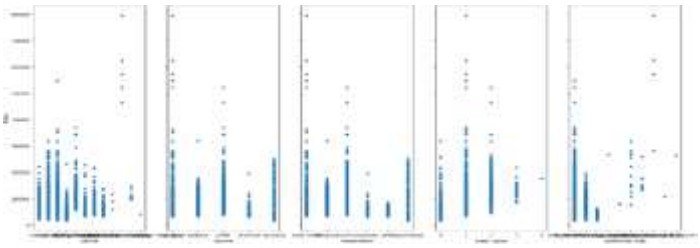


Figure -3: It is a boxplot. In boxplot we can see outlier if it is present. who are paying highest ticket price-based Source. That data is Outlier.

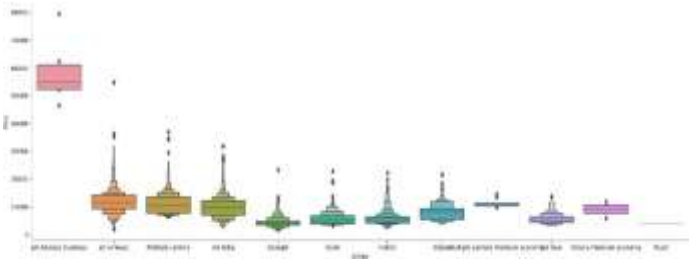


Figure -4: It is a boxplot. In boxplot we can see outlier if it is present. who are paying highest ticket price-based Destination.

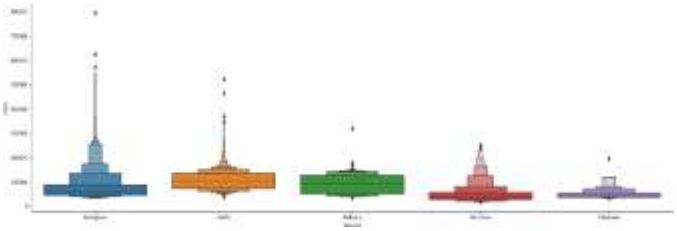
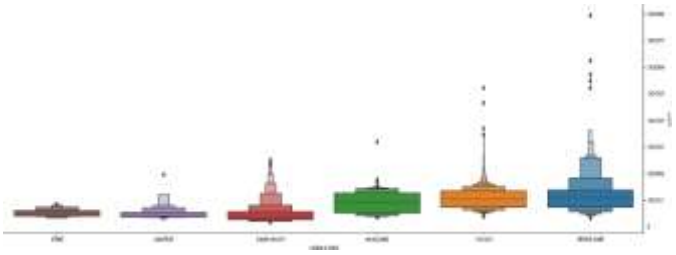


Figure -5: That data is Outlier.



2. METHODOLOGY

The collection of data is the most important aspect of this project. There are various sources of the data on different websites which are used to train the models. Websites give information about the multiple routes, times, airlines and fare. Various sources from APIs to consumer travel websites are available for data scraping. In this section details of the various sources and parameters that are collected are discussed. To implement this data is collected from a website "Makemytrip.com" and python is used for the implementation of the models and collection of the data.

A. Collection of data

The script extracts the information from the website and creates a CSV file as output. This file contains the information with features and its details. Now an important aspect is to select the features that might be needed for the flight prediction algorithm. Output collected from the website contains numerous variables for each flight but not all are required, so only the following feature is considered.

- Origin
- Destination
- Departure Date
- Departure Time
- Arrival Time
- Total Fare
- Airways

B. Cleaning and preparing data All the collected data needed a lot of work so after the collection of data, it is needed to be clean and prepare according to the model requirements. All the unnecessary data is removed like duplicates and null values. In all machine learning this technology, this is the most important and time-consuming step. Various statistical techniques and logic built in python are used to clean and prepare the data. For example, the price was character type, not an integer. from the existing feature. Days to departure can be obtained by calculating the difference between the departure date and the date on which data is taken. This parameter is considered to be within 45 days. Also, the day of departure plays an important role in whether it is holiday or weekday. Intuitively the flights scheduled during weekends have a more price compared to the flights on Wednesday or Thursday. Similarly, time also seems to play an important factor. So, the time is been divided into four categories: Morning, afternoon, evening, night.

IV. MACHINE LEARNING ALGORITHM To develop the model for the flight price prediction, many conventional machine learning algorithms are evaluated. They are as follows: Linear regression, Decision tree, Random Forest Algorithm, K-Nearest neighbors, Multilayer Perceptron, Support Vector Machine (SVM) and Gradient Boosting. All these models are implemented in the scikit learn. To evaluate the performance of this model, certain parameters are considered. They are as follows: R-squared value, Mean Absolute Error (MAE) and Mean Squared Error (MSE). The formulas for these three parameters are as follows

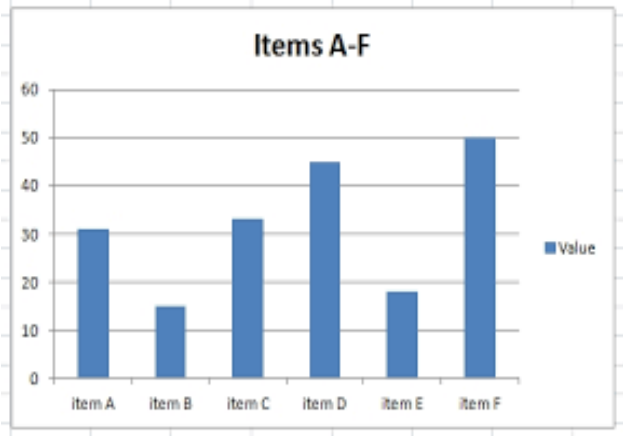
A. Linear Regression is a method of modeling a target value based on predictors that are independent. It is mostly based on the number of independent variables and the relationship between independent and dependent variables. linear regression is a type of analysis where the number of independent variables is one and the relationship between the dependent and independent variables vary linearly. The important concept to understand linear regressions are cost function and Gradient decent. $y(pred) = b0+b1 * x$

B. Decision tree The Decision tree calculation separates the informational collection into small subsets, at a similar same time it creates gradually. The last outcomes are the tree with the decision nodes, what's more, the leaf nodes. A decision hub may have at least two branches. In the beginning, consider the entire informational collection as root. Highlight esteems are wanted to be downright. On the off chance that the qualities are constant then they are discretized before structure the model. Based on characteristic qualities records are dispersed recursively. There are two primary

characteristics in the decision tree calculation. One is Information Gain and another is the Gini index. Information Gain is the proportion of Change in entropy. Higher the entropy more the instructive substance, where the entropy is a proportion of vulnerability of arbitrary variable. Gini Index is a component that measures how frequently an arbitrarily picked.

IJSREM sample template format, define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

Charts



3. CONCLUSIONS

Predicting air fare prices has been extensively studied and various methods and features have been proposed for it to be performed. We gathered the flight price data from the web and clearly showed that it is very much possible and feasible to predict the prices based on some historical data. This report further shows that, ML predictors(models) are a more than satisfactory option to know the air ticket prices. To the level of our understanding, maximum of the preceding research work on the air plane price prediction concept focused on traditional statistical procedures, which have their own obstacles of estimating and prediction. Also, we came to know that proper data and feature extraction and selection are an important part of this process and helped us to come out with some helpful insights. Many features were extracted from the data to make air travel segment easy to understand. We from this project we can get information about the lows and highs of the ticket prices depending on the time of day, the current day or even the weekends. We can also now with full conviction draw a conclusion that if the model implemented in a proper manner can be of great use in saving money of many people by providing them with the information about the air ticket trends and given them an idea about the price which would help them to decide whether to book a ticket right now or in the near future. Also, various ML models are studied and their efficiency and performance are compared so as to get better results. This can also be attributed to the fact that the pricing models used by different airlines are such that to maximize the revenue. Now with help of all this our model can predict the values of the prices with and MSE…

REFERENCES

1. K. Tziridis, T. K. (2017). Airfare Prices Prediction Using Machine Learning Techniques . European Single Processing Confrence, 1017.

2. Rajankar, S. (June 2019). A Survey on Flight Pricing Prediction. International Journal of Engineering Research & Technology (IJERT), 1002.

3. Supriya Rajankar, N. S. (Issue 06, June-2019). A Survey on Flight Pricing Prediction using Machine Learning . International Journal of Engineering Research & Technology (IJERT), 1281.

4. A Survey on Flight Pricing Prediction Using Machine Learning, International Journal of Engineering Research & Technology (IJERT), 2019.

5. Flight Ticket Prediction Using Random Forest Regressor Compared with Decision Tree Regressor, 2023 Eighth International Conference on Science, Technology, Engineering and Mathematics (ICONSTEM), 2023.

6. Integration of Machine Learning Techniques with Existing Systems to Predict Flight Prices, 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2023.

7. Flight Price Prediction Using Machine Learning, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), 2022.

8. Flight Ticket Price Prediction, International Journal of Scientific Research in Engineering and Management (IJSREM), Volume 9, Issue 8, 2025.

9. Flight Fare Prediction Using Machine Learning, The Journal of Computational Science and Engineering, Volume 2, Issue 3, May 2024.

10. Flight Price Prediction Using Machine Learning, Research Gate Publication (PDF), 2024.

11. Flight Price Prediction Using Machine Learning, Paper90.pdf, 2020.

12. Flight Fare Prediction Using Machine Learning, PID39V2I3P11-26.pdf, 2024.

13. Flight Price Prediction, doc.pdf, Research work on dataset analysis and model comparison,2022.

BIOGRAPHIES (Optional not mandatory)

Souvik Dey, currently pursuing B. Tech in Computer Science Engineering from B.P. Poddar Institute of Management and Technology. Right now, in 3rd year.

Mr. Sudipta Kumar Dutta
Currently working as an Assistant Professor in the department of CSE, Ramgarh Engineering College, Jharkhand Ranchi. He has M. Tech degree in computer science from JIS college of Engineering and B. Tech degree in computer science from JIS college of Engineering. His research domain is Artificial Intelligence, Machine learning and Data Mining.