Flight Prices Prediction Model using SparkML

Maria Perez

Department of Information Systems, California State University, Los Angeles

CIS 5560-01 Intro to Big Data Science

e-mail : [mpere110@calstatela.edu](mailto:mpere110@calstatela.edu)

**Abstract:** The dataset provided contains detailed information on flight tickets purchased on Expedia between April 16, 2022, and October 5, 2022. This dataset offers a set of attributes for each ticket that include flight details, pricing information, availability, and travel duration. Specifically, the dataset includes the following key attributes: Flight Details, departure and arrival times, flight durations and aircraft type; Pricing Information, base ticket prices and additional fees or discounts; Availability, number of seats available; Airport Information, airport in the United States identified by their three-character IATA airport codes; and Flight Routes, cities and countries of origin and destination, along with the specific routes for each flight.

By analyzing the dataset using machine learning models, such as Random Forest Tree and Gradient Boosted Tree, we can try to identify the key features that influence ticket pricing. Both models will help us explore complex relationships between various flight characteristics and ticket costs.

# 1. Introduction

Understanding why flight prices fluctuate is essential because prices are influenced by a variety of factors and analyzing these can help both consumers and airlines. Analyzing this data can help consumers by making informed decisions on when to book a flight or assist them in finding the quickest route to their destination at the best possible price. Airlines can benefit from this analysis and use it to maintain competitiveness and maximize profitability.

The dataset used in this project, sourced from Kaggle, contains information on flights between April and October of 2022, covering major U.S. Airports. This data provides insights into factors that influence flight prices, enabling airlines to make more strategic decisions. The primary objective of this project was to develop models capable of accurately predicting flight prices by using key features such as the flight date, departure and destination airports, and base fare.

# 2. Related Work

Predictive modeling has been a subject of analysis in the airline industry. The analysis has focused on achieving key objectives like demand forecasting, price optimization, and improved operational efficiency. These studies have applied machine learning models such as Random Forest Tree and Gradient Boosted Trees models have both been used to identify patterns and make accurate predictions.

# 3. Specifications

This project used Databricks Community Edition to create models that can predict flight prices. Once the models were created, Pyspark CLI through the Hadoop File System was then used to run the final models.

## 3.1 Hardware Specifications

Databricks Community Edition

* Version: 9.1 LTS (Apache Spark 3.1.2, Scala 2.12)
* Memory: 15.3GB
* Cores: 2
* Nodes: 1

Hadoop

* Version: Hadoop 3.3.3/PySpark 3.2.1
* CPU Speed: 1995.312 GHz
* CPU Cores: 8
* Nodes: 5 (2 Master, 2 worker)
* Memory: 806.40GB

## 3.2 Dataset Specifications

Dataset source: https://www.kaggle.com/datasets/dilwong/flightprices

Dataset size: 2.95GB

# 4. Methodology

## 4.1 Data Engineering

Data engineering is the process of collecting, manipulating, and organizing data to make it meaningful and useful. Raw data is often not in the appropriate format for analysis.

In this project, I removed duplicate fields and irrelevant columns. Additionally, missing values were replaced to minimize their impact on the analysis. Data parsing and extraction were also performed on variables such as airport codes and travel durations to ensure the data was structured properly for analysis.

### 4.1.1 Data Cleaning

The first step involves identifying and removing duplicate entries from the dataset to ensure each record is unique. Irrelevant columns, which do not contribute to the predictive model, are also eliminated. Handling missing values is a crucial part of the process, and various imputation techniques, such as mean, median, or mode imputation, are applied based on the nature of the data.

### 4.1.2 Parse and extract from String

Parsing involves analyzing and breaking down a string of text into smaller, meaningful components that can be processed and utilized. After parsing, the relevant sections are extracted for further analysis.

In this project, a function was developed to extract and parse flight duration data, which is provided in the format "PT2H30M". The function converts the duration into total minutes by using regular expressions to identify and capture the hours and minutes from the string.

A screenshot of a computer code

AI-generated content may be incorrect.

## 4.2 Feature Selection

Feature selection is the process of identifying the most important variables that contribute to predicting the target outcomes, such as total fare and remaining seats. For this project, key features selected for analysis include the base flight price, distance travel, travel duration and destination airport. These features will most likely have significant impact on the target variables.

### 4.2.1 Feature Importance

Feature importance involves calculating a score for each input feature in a model, which reflects the contribution of each feature to the model's predictions. A higher score indicates a greater impact on the model’s performance.

For the Random Forest model, Base fare has the highest score of importance followed by distance travel and duration of travel.

Figure 1 Feature Importance for Random Forest

A screenshot of a computer

AI-generated content may be incorrect.

For the Gradient Boosted model, the highest score of importance is also the base fare price of the flight. The base fare has a score of 0.652. The next significant feature has a score of 0.0995, followed by destination airport and total distance travel. These are followed by destination airport and total traveled distance.

Figure 2 Feature Importance for Gradient Boost

A screenshot of a computer

AI-generated content may be incorrect.

Both models have base fare as the feature with the highest feature importance score.

## 4.3 Model Selection

Various machine learning algorithms were explored to identify the best models for predicting total fare.

The algorithms considered include:

Gradient Boosted (GBT): This model can improve prediction accuracy. It can handle complex patterns in the data and improve performance.

Random Forest (RF): This model can handle both numerical and categorical data. It works by constructing multiple decision trees and averaging their predictions. This ensemble approach helps reduce overfitting, making it particularly effective for large datasets.

## 4.4 Model Training and Evaluation

The selected models are trained using the processed dataset, with the data split into training and testing sets to evaluate performance on unseen data. Key metrics like R² and Root Mean Squared Error (RMSE) are used to assess model accuracy.

### 4.4.1 Random Forest

Random Forest is effective because it uses several decision trees to handle both numerical and categorical data, which reduces overfitting and improves prediction accuracy.

### 4.4.2 Gradient Boosted

Gradient Boosted Tree builds decision trees sequentially, each correcting errors made by the previous trees. This technique enhances prediction accuracy and it’s particularly useful for datasets with complex relationships between features, such as the interaction between flight details (like times and durations) and pricing information.

# 5. Results

Anticipating fluctuations in flight prices and seat availability empowers airlines to operate more efficiently and strategically. With accurate forecasts, airlines can implement dynamic pricing models, fine-tune seat allocation, and better align supply with demand—ultimately boosting revenue while enhancing the passenger experience.

## 5.2 R2 and RMSE

R² indicates the proportion of variance in the target variable that the model successfully captures, while RMSE reflects the model’s average prediction error—where lower values signal greater accuracy. Together, these metrics provide a clear picture of model performance, guiding the selection of the most effective approach for forecasting flight fares and seat availability.

Table 1 Results of Random Forest

|  |  |  |
| --- | --- | --- |
|  | Cross Validation | Train Validation Split |
| R2 | 0.85669 | 0.86087 |
| RMSE | 70.27315 | 69.277889 |
| Training Time (sec) | 209.4498 | 99.8695 |

Table 2 Results of Gradient Boosted Trees

|  |  |  |
| --- | --- | --- |
|  | Cross Validation | TrainValidation Split |
| R2 | 0.918024 | 0.918016 |
| RMSE | 53.03872 | 53.04144 |
| Training Time (sec) | 275.0281 | 158.61868 |

Both algorithms showed very similar performance across cross-validation and train validation split, suggesting consistent model behavior regardless of the evaluation strategy. Gradient Boosting outperformed Random Forest in both evaluation methods, achieving a higher R² (~0.92) and lower RMSE (~53.3), indicating better prediction accuracy. However, this came at the cost of longer training times. Gradient Boost was slower than Random Forest.

# 6. Conclusion

Accurately forecasting flight prices and seat availability can transform airline operations by driving smarter pricing decisions, optimizing capacity planning, and supporting long-term strategic goals. Looking ahead, future improvements could involve expanding the range of predictive inputs, such as incorporating real-time data like booking patterns, weather disruptions, or competitor pricing. These dynamic factors can significantly sharpen forecasting precision.

Exploring next-generation machine learning methods, including deep learning architectures, may further elevate model performance. Partnering with industry stakeholders will also be crucial to ensure that these models are not only technically sound but also grounded in operational realities and ready for deployment in real-world settings.

# References

[1] Kaggle: Flight Prices. Retrieved from <https://www.kaggle.com/datasets/dilwong/flightprices>.

[2] S. Boddu, K. S. Reddy, and G. Laks, "EXPLORING PREDICTIVE MODELS FOR AIRFARE FORECASTING: A COMPARATIVE ANALYSIS OF TIME SERIES AND MACHINE LEARNING APPROACHES," Int. Res. J. Mod. Eng. Technol. Sci., vol. 6, no. 3, pp. [5276], Mar. 2024. [www.irjmets.com](https://www.irjmets.com)

[3] A. Lantseva, K. Mukhina, A. Nikishova, S. Ivanov, and K. Knyazkov, "Data-driven Modeling of Airlines Pricing,"

[4] Supermicro, "Data Processing," [Online]. Available: <https://www.supermicro.com/en/glossary/data-processing>. [Accessed: Jan. 08, 2025].