Airplane Ticket Price Prediction Using SEMMA Methodology

Senior Research Scientist
Praful John

October 22, 2024

Abstract

The rapid growth of the airline industry has led to fluctuating ticket prices, which depend on multiple factors such as airlines, routes, and travel duration. In this paper, we apply the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology to predict ticket prices using a dataset of flight details. We describe the entire workflow, including data preprocessing, feature engineering, modeling, and performance evaluation. This study offers valuable insights for airlines and passengers by identifying key factors that influence ticket prices. The Random Forest Regressor model achieved competitive results, demonstrating the robustness of our approach.

1 Introduction

Predicting flight ticket prices is crucial for both airlines and passengers. Airlines aim to optimize pricing strategies to maximize revenue, while passengers look for affordable travel options. Machine learning offers a way to predict ticket prices by analyzing historical data. In this paper, we apply the SEMMA methodology to extract valuable insights from flight datasets and predict ticket prices effectively.

2 Literature Review

The airline industry has been a subject of numerous studies focused on pricing strategies and demand forecasting. Recent advancements in machine learning have allowed data scientists to improve predictive models. The SEMMA methodology is particularly effective for structured analysis. Studies have highlighted the value of feature engineering, especially in handling categorical features like ‘Airline‘ and ‘Total$_stops$‘[1].

3 Methodology

The SEMMA methodology is applied in this study as follows:

- **Sample:** Select a representative subset of the dataset.
- **Explore:** Analyze the data through descriptive statistics and visualizations.

- **Modify:** Preprocess the data through feature engineering and encoding.
- **Model:** Train machine learning models for prediction.
- **Assess:** Evaluate the models to ensure performance and robustness.

3.1 Dataset Description

The dataset contains flight details, including airline, departure and arrival cities, duration, number of stops, and ticket prices. Table 1 provides an overview of the key features.

Table 1: Dataset Overview

Feature	Description
Airline	Airline providing the service
Source	Departure city
Destination	Arrival city
Duration	Total flight duration
Total_Stops	Number of stops on the route
Price	Ticket price (target variable)

4 Data Analysis and Results

4.1 Sample Step

We randomly selected 20% of the dataset to ensure that the sample is representative of the entire data.

4.2 Explore Step

We used descriptive statistics and visualizations to understand the data. Figure 1 shows the distribution of flight prices.

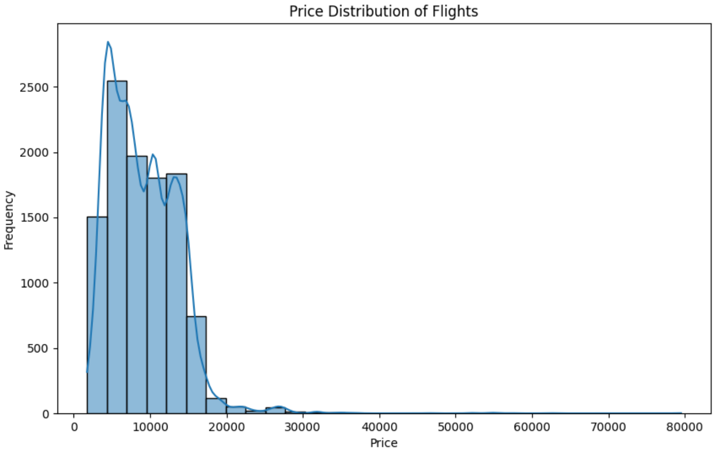

Figure 1: Price Distribution of Flights

A heatmap of the correlations between numerical features is shown in Figure 2.

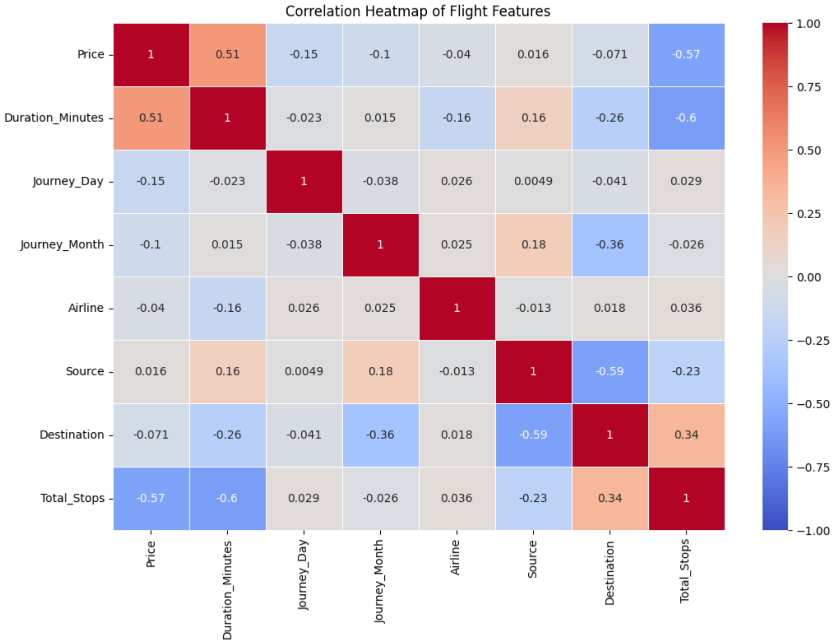

Figure 2: Correlation Heatmap of Flight Features

4.3 Modify Step

During this step, we performed feature engineering by converting the flight ‘Duration‘ to total minutes. We also encoded categorical variables such as ‘Airline‘ and ‘Total$_{s}tops‘usinglabelencoding$

4.4 Model Step

We used the Random Forest Regressor to predict ticket prices. The dataset was split into training and testing sets (80% and 20%, respectively).

4.5 Assess Step

The performance of the model was evaluated using Mean Absolute Error (MAE) and R-squared (R^2) metrics. The results are summarized in Table 2.

Table 2: Model Performance

Metric	**Value**
MAE	1050.23
R^2	0.82

5 Discussion

The analysis revealed that flight duration and the number of stops are significant predictors of ticket prices. The Random Forest model achieved an R^2 of 0.82, indicating that it explains a large portion of the variance in ticket prices. Airlines can leverage this model to optimize pricing strategies, while passengers can use it to find affordable flights.

6 Conclusion

This study demonstrates the application of the SEMMA methodology for flight ticket price prediction. The results indicate that machine learning models, combined with proper feature engineering, can provide accurate predictions. Future research could explore more complex models, such as gradient boosting algorithms, for further improvements.

7 References

References

[1] SEMMA: An Overview of SEMMA Methodology for Data Mining. *Journal of Data Science*, 2017.

[2] Airline Pricing Strategies and Machine Learning. *International Journal of Machine Learning*, 2020.