Flight Prices Prediction

Authors

Najd Alsabi, Asma Alshilash, Joud Almutairi, Leena Almusharraf, Lujain Alharbi, Hissah Alhano, Khulood Alyahya
Emails:
443200578@student.ksu.edu.sa , 443200439@student.ksu.edu.sa,
443200544@student.ksu.edu.sa , 443200563@student.ksu.edu.sa ,
443200811@student.ksu.edu.sa , 443200617@student.ksu.edu.sa ,
Kalyahy1@ksu.edu.sa

Abstract

This study investigates the complex dynamics of airline ticket pricing in Saudi Arabia, where rapid growth in air travel has heightened demand and pricing variability. Utilizing a dataset of 505,504 flights collected from Google Flights through web scraping, the research focuses on routes between Riyadh and 21 other cities, with data gathered at one month, one week, and one day prior to departure to analyze price fluctuations. The findings reveal significant adjustments in pricing strategies by airlines in response to temporal factors and demand patterns. Additionally, the study compares pricing trends in Saudi Arabia with those in the United States, identifying key regional differences in pricing dynamics, which highlight the unique challenges and opportunities faced by travelers in each market. To address these insights, a predictive model is developed to estimate flight ticket prices, providing users with valuable tools for informed decision-making when booking flights. This research offers critical insights for consumers and industry stakeholders alike, facilitating a better understanding of the factors influencing airline ticket pricing.

Keywords

Flight Fare, Booking Timing, Departure Date Impact, Machine Learning Algorithms, Search Date Influence, Data Analysis, Predictions

1. Introduction

Background

Airline ticket pricing is a dynamic and complex process influenced by numerous factors such as booking timing, travel routes, demand fluctuations, and airline-specific pricing strategies. For travelers, the unpredictability of ticket prices often creates challenges in determining the best times to book flights, particularly when ticket prices vary significantly as the departure date approaches. This complexity is heightened in regions like Saudi Arabia, where rapid growth in air travel has led to increased demand, competition, and variability in pricing. Understanding how these factors impact flight prices can provide valuable insights for both consumers and industry stakeholders.

In this study, we address these complexities by examining a dataset of 505,504 flights collected from Google Flights through web scraping. The dataset focuses on flights between Riyadh and 21 other cities within Saudi Arabia, providing a comprehensive view of travel pricing trends. To observe fluctuations, data was collected at intervals of one month, one week, and one day before departure. This approach allows for an in-depth analysis of how flight prices shift as the departure date nears, offering insights into how airlines adjust prices in response to temporal factors and demand. By comparing this primary dataset with secondary data, this research aims to provide a nuanced understanding of price variation and the factors contributing to these differences.

Objectives

The primary objective of this research is to analyze the key factors influencing flight prices within Saudi Arabia, focusing on how booking timing (one month, one week, and one day before departure) affects ticket costs. The study aims to identify patterns specific to routes between Riyadh and selected cities, exploring how variables like destination, demand, and airline company contribute to price changes. Additionally, we will compare our collected primary dataset with a secondary dataset sourced from Kaggle, which focuses on flights in the United States. This comparison will help uncover notable differences in pricing dynamics between the two regions, providing a broader context for how flight prices may vary under different conditions or data sources. To extend these insights, we will develop a predictive model capable of estimating flight prices, offering consumers a valuable tool for making informed booking decisions.

2. Related Work

Literature Review: We summarized some studies related to our project to gain a deeper understanding, and most of the studies focused on predictive modeling for airfare and ticket prices, highlighting methods that enhance model performance and accuracy. Principal Component Analysis (PCA) reduces dimensionality while retaining essential information, which speeds up processing and prevents overfitting. Outlier removal reduces noise by eliminating extreme values, enhancing model robustness, and resampling addresses data imbalance to minimize bias toward dominant classes. Model selection in these studies spans linear models like Linear and Ridge Regression, suitable for baseline and interpretable predictions, to tree-based models like Decision Trees, Random Forest, and XGBoost, which capture complex, non-linear patterns. Ensemble models, such as AdaBoost and Random Forest, combine predictions across multiple models to improve accuracy, while deep learning models (e.g., MobileNetV3, VGG19) handle complex data patterns well, making them effective for real-time applications in large datasets. Quantum Machine Learning models (e.g., QSVM) show potential for high accuracy in structured data but demand significant computational resources. Performance metrics in these studies include Mean Squared Error (MSE) and Mean Absolute Error (MAE) to assess error magnitudes, R-squared for fit quality, and Root Mean Square Error (RMSE), providing interpretable error values in the same units as the target. To ensure unbiased assessment, dataset splitting for training and testing is employed, while hyperparameter tuning using techniques like Randomized Search CV optimizes configurations, improving model accuracy and generalizability. Together, these methods improve model adaptability, accuracy, and stability—key for predicting dynamic airfare and ticket prices.

The primary gaps include limited inclusion of real-time variables, narrow data scopes (e.g., limited routes or regions), and dependency on specific datasets that may miss broader travel patterns or post-pandemic trends. Runtime constraints often restrict the use of complex models like RandomForest and SVR, impacting predictive accuracy. Additionally, dimensionality reduction techniques like PCA may inadvertently remove subtle data insights. For airfare prediction, geographical limitations (e.g., European airlines only) and quarterly data aggregation hinder capturing short-term price fluctuations, while reliance on external economic factors makes models vulnerable to volatility. Future research could address these gaps by integrating real-time data, expanding datasets to include diverse regions and timeframes, and exploring models that balance interpretability with high-dimensional data handling.

Contribution: we aim to advance understanding of airline ticket pricing within Saudi Arabia by examining key factors that influence price fluctuations, particularly focusing on booking timing. Our Approach is to identify unique pricing patterns on routes between Riyadh and selected cities, analyzing how variables like destination, demand, and airline policies impact costs. Additionally, by comparing our primary dataset with secondary data, we aim to uncover differences in pricing dynamics, ultimately developing a predictive model that empowers travelers to make informed booking decisions.

3. Data Collection and Preprocessing

Data Sources:

Primary Dataset:

- **Dataset Used:** The primary dataset for this study comprises detailed flight data focusing on key aspects such as flight duration, ticket prices, and airline frequency. This dataset was specifically collected for this project as it captures a broad set of information not readily available in public datasets.

- **Origin and Structure:** The data was gathered from Google Flights via web scraping, providing comprehensive information on flight schedules, costs, airlines, departure and arrival cities, and seasonal trends.

- **Size and Features:** This proprietary dataset includes essential variables such as flight duration, type of airline, and ticket price, offering a complete picture of the market. The structured data was customized to meet the project's needs, ensuring robust insights.

Secondary Dataset:

- **Dataset Used:** To complement the primary data, a secondary dataset was collected, providing additional insights into flight trends with similar key features. This dataset serves as a comparative source, enabling broader context alongside the primary data.

- **Origin and Structure:** the secondary data was sourced externally from Kaggle, aiming to capture different regional or seasonal patterns.

- **Size and Features:** With 317,260 rows and 11 columns, the secondary dataset covers similar features—such as flight duration, airline type, and ticket price—ensuring that it aligns well with the primary data for straightforward comparison.

Data Collection Process:

Primary Dataset Collection:

- **Collection Tools:** The primary dataset was collected through web scraping to gather information from Google Flight after attempts to collect data via SkyScanner and SerpAPI were unsuccessful due to limitations.

- **Techniques:** Web scraping was primarily used, capturing flight data over different periods to observe seasonal trends and analyze factors affecting flight prices and durations.

Secondary Dataset Collection:

- **Collection Tools:** The secondary dataset was sourced from Kaggle, aiming to capture different regional or seasonal patterns. Although the specific methods for data collection remain unclear.

- **Techniques:** This dataset was structured to match the primary dataset, allowing for side-by-side analysis to validate and expand upon the insights from the primary data.

Data Preprocessing:

Primary Dataset Preprocessing steps:

- **Handling Missing Values**: Steps were taken to address missing data entries to ensure dataset completeness and accuracy.

- **Time Formatting:** The time data was converted into a uniform format for consistent comparison.

- **Data Cleaning:** The dataset underwent a cleaning process to remove inconsistencies, standardize formats, and prepare it for further analysis.

- **Duplicate Removal:** Duplicate records were detected and removed to maintain the integrity and accuracy of the dataset.

Secondary Dataset Preprocessing steps:

- **Handling Missing Values:** Missing values were managed, and fields like the Route column were removed to streamline analysis.

- **Data Transformation:** The Price column was converted from USD to SAR for consistency with the primary dataset, enabling easy comparison.

- **Time and Date Formatting:** Columns for Departure Date, Departure Time, and Arrival Time were reformatted to a 24-hour standard, and extraneous symbols were removed for consistency.

- **Duplicate Removal:** Duplicate records were identified and eliminated, ensuring accuracy in the dataset.

6. Discussion

Regional Bias:

Our data primarily focuses on specific airlines, regions (mainly Saudi Arabia), and routes, with a significant presence of Saudia airline. Consequently, the findings predominantly reflect regional trends within Saudi Arabia rather than global patterns, which limits their generalizability to other markets or regions with different travel behaviors and airline options. Additionally, our secondary data is sourced exclusively from the USA, further constraining the applicability of the results, as insights derived from these two regions may not be universally applicable across other markets with diverse travel dynamics and airline preferences.

Time Constraints:

While the dataset spans a full year (24 September 2024 to 24 September 2025), capturing all seasons, specific events or short-term trends—such as a particular destination becoming trendy—may still impact prices or schedules. As such, findings may not fully reflect changes in flight demand or economic factors that affect pricing, which could result in misleading prices for some destinations during this period.

Limited Feature Scope:

Missing factors like passenger demand, economic conditions, and flight capacity can significantly affect flight prices and durations. High demand raises prices, economic shifts (like fuel costs) impact airline expenses, and limited seat availability increases ticket prices. Without these, observed relationships between airline, price, and duration may be incomplete.

Restricted City Pairs:

The analysis was limited to flights departing from or arriving to Riyadh (RUH). For example, routes include Riyadh to Jeddah (RUH to JED) and Dammam (RUH to DMM), and vice versa. No data was collected for routes between other cities, such as Jeddah to Dammam (JED to DMM). This limitation means the findings reflect only Riyadh-related travel patterns, which may not represent trends for other city pairs.

Correlation Analysis:

The weak correlations observed between variables (between flight duration and price) may indicate that the features alone are not sufficient to predict prices accurately.

Outliers and Skewed Distributions:

The presence of high-priced outliers (flights over 25000 SAR) skews the data. This may affect the average-based measures and certain analyses. While these outliers provide valuable insights into premium pricing, they can also distort general pricing trends in the dataset.

Limited Timeframe in Secondary Data:

Our secondary dataset, sourced exclusively from the USA, covers only three non-consecutive months (July, September, and December 2024), which are primarily vacation or holiday seasons. This limited timeframe restricts its ability to represent year-round pricing trends, as it lacks data from non-peak periods. Consequently, the analysis may not fully

capture seasonal fluctuations in flight ticket prices, impacting the accuracy and generalizability of insights across the entire year.

7. Reference

Studies:

- Machine learning modeling for time series problem: Predicting flight ticket prices
- Flight Fare Prediction Using Machine Learning
- A Holistic Approach on Airfare Price Prediction Using Machine Learning Techniques
- Using Different Machine Learning Algorithms to Predict the Prices of Flight Tickets
- A Framework for Airfare Price Prediction: A Machine Learning Approach
- Prediction of Flight-fare using machine learning