Prediction of Flight Price

Group Project Progress | EAS 508 Statistical Learning & Data Mining – I | Spring 2024

Team number: 07

Team: Rakesh Nandan, Shivangi Rai, Santoshi Reddy Chintala, Saahithi Chippa

Abstract

This study analyzes historical flight data to identify price-influencing factors and establish categorical price tiers. Utilizing exploratory data analysis, correlations and patterns impacting pricing are examined. A categorization system classifies ticket prices into low, medium, or high categories based on deviations from median prices per source-destination pair. Comparing linear regression and random forest models using MSE, RMSE, and MAE, supported by cross-validation, enhances price forecasting accuracy in the airline industry.

Introduction

The prediction of flight prices is crucial in the airline industry, influencing both consumer decisions and revenue management strategies. In this project, we aim to develop a predictive model for flight prices using machine learning techniques. By analyzing a comprehensive dataset containing information about various flight attributes such as departure and arrival times, airlines, source and destination, total stops, duration and prices. Our goal is to build a model capable of accurately estimating flight prices. Through a systematic approach involving data preprocessing, exploratory data analysis (EDA), feature selection, model selection, training, and evaluation, we strive to identify the most relevant features and develop robust regression models for predicting flight prices. This report presents the methodology, findings, and insights derived from the analysis, highlighting the significance of predictive modeling in optimizing pricing strategies and enhancing the overall customer experience in the airline industry.

Literature Survey

We conducted a comprehensive literature survey to understand existing research in predicting flight prices. Etzioni et al. (2003) [1] explore the utilization of data mining techniques to minimize airfare prices, with a goal of aiding passengers in making cost-effective ticket purchases. Abdella (2019) [2] delves into methodologies for predicting airline ticket prices and demand, encompassing factors such as seasonality, route popularity, and economic indicators, crucial for accurate predictive modeling of flight costs. Balasubramanian's [3] work contributes to understanding flight operations by developing predictive models using binary classification and regression trees to improve flight on-time performance. This approach aims to minimize aircraft delays and increase operational efficiency by accurately predicting arrival flight timings.

Research employing gradient boosted decision trees and historical on-time performance data seeks to predict flight delays in the United States, offering potential solutions to the issue of unpredictable delays stemming from various causes [5]. If successful, this predictive model could aid airlines, airports, and air traffic management in anticipating and managing flight delays, thereby enhancing operational efficiency and customer satisfaction. Future research directions include integrating real-time weather and air traffic information to enhance predictive capabilities and address limitations in forecasting unusual events such as severe weather [6].

Additional investigations include Tziridis' study on forecasting airfare prices using flight-specific characteristics, Wang's integration of macroeconomic information to predict average ticket prices across

different market segments, and proposals by Groves and Gini for estimating probable costs based on consumer preferences [10,11,12]. Current practices in predicting ticket prices rely on factors like journey date, booking time, and inflation but lack accuracy. To enhance accuracy, novel approaches such as the Novel XGBRegressor Optimizer algorithm and comparisons with the Extra Tree Regression algorithm have been introduced [7,8]. Implementation timelines depend on factors such as data availability, model development complexity, and deployment infrastructure, but with dedicated resources, implementation can be achieved within a reasonable timeframe [Heilmeier's q8].

Key findings include the application of data mining techniques, machine learning algorithms, and predictive modeling methods to forecast airline ticket prices. The survey provided insights into methodologies, challenges, and potential innovations in this domain.

Project Overview:

1. **Data Collection and Exploration**: We collected historical flight data from relevant sources. The dataset consists of around 10,700 records showcasing type of airlines, Source, destination, time and departure of travel, route, date and price based on the different predictors mentioned. We have examined the data to understand the datatypes, missing values and other summary statistics.
2. **Data Preprocessing**: We have performed preprocessing steps to clean and prepare the dataset for analysis. This includes:
    a. Dropping missing values
    b. Removing duplicate rows and outliers
    c. Extracting features from date column
    d. Handling 'total_stops' column of string to integers for better interpretation of data
    e. Standardizing values in different columns by merging similar values.
    f. Creating dummy variables for categorical variables
    g. Dropping columns that are not required for modeling.

3. **Exploratory Data Analysis (EDA):** EDA was conducted to gain insights into the dataset's structure and distribution. We explored correlations and patterns influencing flight prices, examining factors such as journey date, time of booking, route popularity, and economic indicators. We have also performed univariate analysis (examining individual variables), bivariate analysis (exploring relationships between pairs of variables). Visualizations such as box plots and bar plots are used to illustrate trends and relationships in the data.

4. **Feature Selection:** We have performed regression analysis to rank the importance of features based on their contribution to predicting flight prices. The top ranked features are used for further analysis and modeling.

5. **Data Modeling**: We implemented various machine learning algorithms, including linear regression and random forest models, to predict flight ticket prices. Model performance was evaluated using metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Cross-validation techniques were employed to assess the models' robustness and generalization ability.

6. **Model Training**: This step involves fitting the above-mentioned regression models and random forest models to the training data (80% of the dataset) to learn the underlying patterns and relationships. Each regression model is trained using the training data and the model's performance is evaluated.

7. **Pending Tasks**: The project is currently in the data interpretation and evaluation phase, where further analysis and validation will be conducted to assess the impact of different factors on model performance.
    a. We tend to understand the relationships between input features and the target variable learned by the trained models. Interpretability may vary depending on the complexity of the model. Linear regression models provide coefficients that indicate the strength and direction of the relationships, while decision trees offer insights into feature importance and splitting criteria.
    b. Model evaluation involves assessing the performance of the trained regression models using appropriate evaluation metrics. Common evaluation metrics include R-squared (R2) score, adjusted R-squared score, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). These metrics help us in determining how well the model generalizes to unseen data and further refining them to improve accuracy and reliability.

Challenges:

- **Data Quality**: Ensuring the accuracy and completeness of the dataset posed challenges during preprocessing, requiring careful handling of missing values and outliers.
- **Model Selection**: Choosing the most suitable machine learning algorithms and techniques for price forecasting involved experimentation and evaluation to identify optimal solutions.

Next Steps:

- Data Evaluation: Further analysis and validation of the models will be performed to assess their performance and accuracy.
- Model Refinement: Refining the models based on evaluation results and incorporating feedback to improve predictive accuracy.
- Documentation and Reporting: Documenting the project findings, methodologies, and results for presentation and dissemination.

All team members have contributed a similar amount of effort

Conclusion: The project has made significant progress in data collection, preprocessing, exploratory analysis, and model development for predicting flight ticket prices. The analysis reveals key features that significantly influence flight prices, such as departure and arrival times, total stops, airline, and duration of the flight. Understanding these factors is essential for airlines to optimize pricing strategies and enhance revenue management. The developed regression models demonstrate varying degrees of predictive accuracy in estimating flight prices. Evaluation metrics such as R-squared (R2) score, mean

absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) provide insights into the models' performance and their ability to generalize to unseen data. The project provides valuable insights into the factors driving flight prices and their respective impacts. Airlines can leverage these insights to make informed decisions regarding pricing strategies, route optimization, scheduling, and customer segmentation, ultimately enhancing competitiveness and profitability. With ongoing data evaluation and model refinement, we aim to deliver accurate predictive models that contribute valuable insights to the aviation industry.

References:

1. *To buy or not to buy: mining airfare data to minimize ticket purchase price Authors: Oren Etzioni, Rattapoom Tuchinda, Craig A. Knoblock, Alexander Yates,KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. August 2003. Pages 119–128.*
2. *Airline ticket price and demand prediction: A survey. Citation: DataJournal of King Saud University - Computer and Information Sciences, ISSN: 1319-1578, Vol: 33, Issue: 4, Page: 375-391. Publication Year2021*
3. *B. Thiagarajan, L. Srinivasan, A. V. Sharma, D. Sreekanthan and V. Vijayaraghavan, "A machine learning approach for prediction of on-time performance of flights," 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), St. Petersburg, FL, USA, 2017, pp. 1-6, doi: 10.1109/DASC.2017.8102138.*
4. *Data-driven flight time prediction for arrival aircraft within the terminal area – The Insitute of Engineering & Technology, IET Intelligent Transport Systems, Volume 16 – Authors: Junfeng Zhang, Zihan Peng, Chunwei Yang & Bin Wang.*
5. *Prediction of Flight Delay Using Gradient Boosted Decision Tree – Research Gate - Authors: Jitendra Kumar Jaiswal & Rita Samikannu*
6. *A Review on Flight Delay Prediction - Transport Reviews,Volume 41, 2021 - Issue 4 - Authors: Alice Sternberg, Jorge Soares, Diego Carvalho & Eduardo Ogasawara*
7. *Improve the Accuracy for Flight Ticket Prediction using XGBRegressor Optimizer in Comparison with Extra TreeRegressor Performance. - published in 2023 international conference (IC3I) - Author: Saatwik Kumar G.V, Jaisharma K.*
8. *Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not - Published: 19 July 2022 – Authors: This work is distributed under the Creative Commons Attribution.*
9. *Airline Prices Analysis and Prediction Using Decision Tree Regressor by Neeraj Joshi, Gaurav Singh, Saurav Kumar, Rachna Jain & Preeti Nagrath. May 2020. Communications in Computer and Information Science.*
10. *Tziridis, K., Kalampokas, T., Papakostas, G. A., & Diamantaras, K. I. Airfare prices prediction using machine learning techniques. 25th European Signal Processing Conference (EUSIPCO), IEEE, 1036-1039, 2017.*
11. *Wang, T., Pouyanfar, S., Tian, H., Tao, Y., Alonso, M., Luis, S., & Chen, S. C. "A framework for airfare price prediction: A machine learning approach". 20th International Conference on Information Reuse and Integration for Data Science (IRI), IEEE, pp. 200-207, 2019.*
12. *Groves, W., & Gini, M. "A regression model for predicting optimal purchase timing for airline tickets," Technical Report 11-025, University of Minnesota, Minneapolis, 2011.*