Project Report

Data Mining Outline Group 4

Benedetta Fiorillo, Chien-Yun Lin, Javier Miranda,
Jieun Park, Tatiana Nachev, Yujeong Kim

March 23, 2025

Submitted to
Data Mining Teaching Team
Prof. Dr. Sven Hertling
University of Mannheim

Contents

1. Introduction

In this project, our primary goal is to analyze and predict flight ticket prices using historical data. We aim to investigate how various variables influence the price of a flight. Understanding price trends and identifying key influencing elements can assist both consumers and airlines in making more informed decisions. Specifically, the analysis will be focused on identifying patterns in flight pricing over time, evaluating the impact of different flight characteristics and ultimately developing predictive models capable of estimating flight prices based on relevant features.

To perform our analysis, we will use the ”Flight Prices” dataset (Kaggle 2023), which is publicly available in CSV format. The dataset is accessible through Kaggle and contains detailed records of flight tickets found on the online travel agency Expedia. This dataset covers a period from April 16th, 2022, to October 5th, 2022, spanning nearly six months, and includes flights between various major airports in the United States. The data will be downloaded from Kaggle and imported into a Python-based environment for processing and analysis. This dataset includes information such as departure and arrival airports, travel duration, ticket prices, seat availability, and details about the airline and aircraft, providing a comprehensive foundation for our study. A list of the features included in the dataset is available in Table A.1. For this study, the base fare of each flight search will be used as the target variable, predicting the flight price.

2. Preprocessing

The dataset for this analysis contains 82.138.753 rows, detailing flight information between 16 airports across the United States from April to October 2022. Given the extensive size of the dataset, the analysis will be focused exclusively on flights departing from a single airport. Los Angeles International Airport (LAX) has been selected, as it comprises 8.073.281 rows, representing approximately 9,83% of the total data. To enhance manageability while preserving representativeness, the dataset will be limited to flights from the summer months (June to September). This period aligns with peak air travel in United States as evidenced by the Transportation Security Administration (TSA), which reported that July 2022 recorded over 70 million passengers, marking the highest travel volume of the year (Transportation Security Administration 2022). A stratified sampling approach based on the distribution of flights across these months will then be employed to systematically reduce the dataset to 150.000 rows, ensuring a representative selection of data for analysis.

Having the reduced dataset, we will continue with the preprocessing which is categorized into four main operations: removal, conversion, aggregation of multi-leg flights, and addition of features.

1. Removal
    - Eliminating records with missing values to maintain data integrity.
    - Validating feature values and removing invalid entries, such as incorrect date formats or string values in numerical fields.

- Removing redundant features that store identical information in different formats (e.g., Departure Time in Epoch Seconds and Departure Time Raw: the same information but in different timestamp format).
- Dropping features with constant values that do not contribute to the learning process (e.g., is Refundable).
- Identifying and removing duplicate entries.
- Detecting potential outliers by using boxplots and excluding them.

2. Conversion
    - Categorical to numerical conversion: Categorical attributes are converted into nominal attributes using One-Hot Encoding.
    - Normalization of numerical features: Standardization is applied using StandardScaler or MinMaxScaler for attributes such as travel distance and duration.
    - Date format standardization: Date values are parsed and unified into a consistent format.

3. Aggregation of multi-leg flights

    The dataset contains features related to multi-leg flights, where legs are separated by ||. To condense this information for more efficient analysis, we aggregate the following:

    - Flight Times: Combine the departure and the arrival time of legs to capture the entire flight duration including transfer time.
    - Airports and Legs: Aggregate the codes of departure and arrival airports to count legs and identify intermediate airports.
    - Airline Diversity: From the code of airlines, derive a feature indicating whether multiple airlines are involved and extract the airline for each leg.
    - Aircraft Type: Extract the type of aircraft for each leg.
    - Duration and Distance: Sum the duration of all legs as well as the distance to get the total flight time and distance.
    - Cabin Class: From the cabin code of the flights, create a feature to indicate if only economy class was selected.

4. Addition
    - Adding date-related features that may influence ticket prices, including day of the week (integer), month of the year (integer), holiday (boolean), and days remaining until departure (integer), among others.

3. Model Choices and Evaluation Methods

3.1. Model Choices

Our target variable will be the base fare, representing the flight ticket price before applying the US taxes and our modeling strategy involves comparing several methods for prediction.

- As a starting point, we use a simple baseline with the **DummyRegressor** set to the "mean" strategy, for obtaining a minimum performance threshold against which more sophisticated models can be evaluated.

- Our first advanced model is **K-Nearest Neighbors (KNN) Regression** with distance-based weights. In our approach, we will use Euclidean distance as our distance metric.

  Hyperparameters to tune: the number of neighbors (k),

- Next, we incorporate **Linear Regression**, where we apply both Lasso and Ridge regularization. We plan to compare the performance of Lasso and Ridge regression to determine which regularization method best suits our dataset.

  Hyperparameters to tune: the regularization strength (λ), which controls the penalty imposed on the coefficient sizes.

- Our third model is a **Regression Tree**. Regression trees partition the feature space into regions and make predictions by averaging the target values within each region. In our approach, splits will be selected by maximizing the MSE reduction.

  Hyperparameters to tune: the maximum depth of the tree and the minimum number of samples per leaf.

- Finally, we utilize an **Artificial Neural Network (ANN)** for regression, which is well-suited for modeling complex, non-linear interactions among features. In our approach, we will use the ReLU activation function in the hidden layers and a linear activation function in the output layer to generate a continuous price prediction.

  Hyperparameters to tune: the network architecture (i.e., the number of hidden layers and neurons per layer), the learning rate, the batch size, and the number of training epochs.

We will optimize each model's performance through hyperparameter optimization, ensuring that we systematically identify the best settings for every algorithm. For model selection, we plan to choose between using a dedicated validation set or employing cross-validation techniques, after analyzing the size of our final dataset.

3.2. Evaluation Methods

These evaluation methods will be implemented to estimate how our model explains the actual flight prices.

- Mean Absolute Error (MAE): Measures the average of the absolute differences between the actual and predicted base fare of flights. It is robust when flight ticket prices are extremely small.
- Root Mean Squared Error (RMSE): Gives an insight of how far off the predictions are in the same units as the target variable (Base fare). Since it penalizes large prediction errors more heavily, it helps to prevent a greater impact on large errors. This is crucial when dealing with price variances across different routes, dates, time, distance, or airlines.
- Adjusted R-squared: Helps prevent overfitting by penalizing the addition of unnecessary independent variables. Assessing the proportion of variance in flight prices can be explained by computing adjusted R-squared.

4. Expected Results

We expect our results to include a reliable regression model capable of accurately estimating flight ticket prices based on input features. In addition, the analysis will present meaningful insights into the factors that most strongly impact price variations. Visualizations, and trend analyses will support these findings.

Bibliography

[Kaggle 2023] Wong, D. (2023). 2022 Expedia Flight Prices Dataset. Available at: https://www.kaggle.com/datasets/dilwong/flightprices/data

[Transportation Security Administration 2022] 2022 TSA checkpoint travel numbers. Available at: https://www.tsa.gov/travel/passenger-volumes/2022

A. Appendix

Table A.1.: Flight Price Dataset Features (Self-made)

Information area	**Features in the data set**
Flight	Flight date, Starting airport, Destination airport, Travel duration, Number of elapsed days, Boolean indicating if a ticket is basic economy, Boolean indicating if a ticket is refundable, Boolean indicating if a ticket is non-stop, Number of remaining seats, Travel distance, Departure time, Arrival time.
Multi legs	Departure time for each leg, Arrival for each leg, Arrival airports in each leg, Departure airports in each leg, Airline name for each leg, Airline code for each leg, Airplane used for each leg, Duration for each leg, Distance for each leg, Cabin used in each leg.
Price	Base fare, Total fare (after taxes and fees).
Flight search	Search date.

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT4o	Rephrasing	Throughout	++

B. Fiorillo C. Lin J. Miranda J. Park

T. Nachev G. Kim

Unterschrift
Mannheim, den 23. 03 2025