

Machine learning Web Application Report

Project Title: Civil Aviation



September 18, 2024

888 – Peaky blinder

Xuan Tuan Minh Nguyen – Ba Viet Anh (Henry) Nguyen – Trong Dat Hoang

Tutor: Ningran Li

# Introduction

In an era of increasingly complex air travel, understanding and predicting flight prices has become crucial for both consumers and airlines. Our machine learning project aims to address this challenge by developing predictive models that can accurately estimate flight prices based on various factors. The primary motivation behind this project is to empower travelers with insights that can help them make informed decisions about their air travel purchases, potentially saving money and reducing the stress associated with ticket booking.  
  
Our intended users are primarily budget-conscious travelers who want to optimize their travel expenses. These users range from occasional vacationers to frequent business travelers who need to balance cost with convenience. By providing them with a tool that can predict flight prices, we aim to help them plan their trips more effectively, choosing the most cost-effective options without compromising on their travel needs.

# Problem Framing

The challenge of predicting flight prices is apparently multifaceted and complex. Airlines use dynamic pricing strategies that take into account numerous variables, including demand, seasonality, fuel costs, and competitor pricing. This complexity makes it difficult for consumers to anticipate price fluctuations and make informed purchasing decisions.

Existing solutions, such as fare comparison websites, provide current prices but lack predictive capabilities. They don't offer insights into future price trends or the factors influencing these changes. This limitation often leads to suboptimal purchasing decisions, where travelers either buy too early at higher prices or wait too long and miss out on better deals.

A machine learning approach is particularly suitable for this problem due to its ability to:

1. Handle large volumes of data with multiple variables
2. Identify complex patterns and relationships that may not be apparent through traditional analysis
3. Continuously learn and adapt to changing market conditions
4. Provide probabilistic predictions that can account for the inherent uncertainty in price fluctuations

By leveraging machine learning, we can create models that not only predict prices but also provide insights into the factors driving these predictions, offering a more comprehensive solution to the challenge of flight price optimization.

# Data Collection

## Source and Method

For our analysis of flight pricing in German civil aviation, we utilized the "German Domestic Air Fares" dataset, which was downloaded from Mendeley Data, a reputable open-access data repository. This section details our data collection process, including sources, methods, criteria, and challenges encountered.

A screenshot of a computer

Description automatically generated

The primary dataset was obtained from Mendeley Data, published on January 6, 2021, by contributor Frederick F and associated with Zeppelin Universität. The data was originally collected through web scraping techniques, gathering information on ticket prices for 84 German domestic flight connections over a 6-month period.

## Data Collection Criteria

In selecting this dataset, we applied the following criteria to ensure its relevance and sufficiency for our analysis:

1. Relevance to German civil aviation: The dataset specifically focuses on domestic flights within Germany, aligning perfectly with our project scope.
2. Comprehensiveness: With 63,000 data points covering 84 routes, the dataset provides a broad and representative sample of the German domestic air travel market.
3. Recency: Published in 2021, the data reflects relatively recent pricing trends, crucial for developing a current predictive model.
4. Feature richness: The dataset includes key variables such as departure/arrival cities, dates, times, airlines, and prices, providing a solid foundation for our analysis.
5. Temporal span: Covering a 6-month period, the data allows for analysis of seasonal trends and temporal variations in pricing.
6. Credibility: Association with Zeppelin Universität, a recognized institution, lends academic credibility to the data.

## Challenges and Solutions

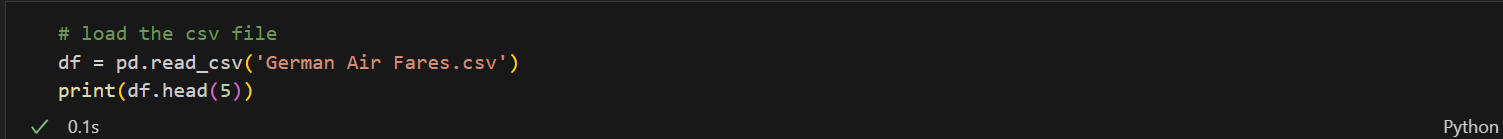
We have faced some challenges when collecting the data. However, we discussed together and addressed the problems:

1. Limited additional data sources: We initially aimed to supplement the primary dataset with additional information such as weather data or economic indicators. However, finding relevant, open-source data that aligned with our specific routes and timeframe proved challenging. Solution: We focused on maximizing the value of our primary dataset through extensive feature engineering, creating derived variables to capture additional insights.
2. Data volume management: The dataset, while comprehensive, was large (7.15 MB) and required efficient handling to ensure smooth processing. Solution: We optimized our data loading and processing pipeline using pandas and NumPy libraries in Python, ensuring efficient memory usage and processing speed.
3. Potential bias in web-scraped data: The original web scraping method used to collect the data could have introduced biases, such as focusing on specific websites or times of day. Solution: We carefully examined the data distribution and collection patterns to identify any apparent biases. Where necessary, we implemented statistical techniques to mitigate these effects.
4. Ethical considerations: Using web-scraped data raised questions about the ethical implications of our analysis. Solution: We ensured our use of the data complied with Mendeley Data's terms of service and focused our analysis on aggregate trends rather than individual flight details.

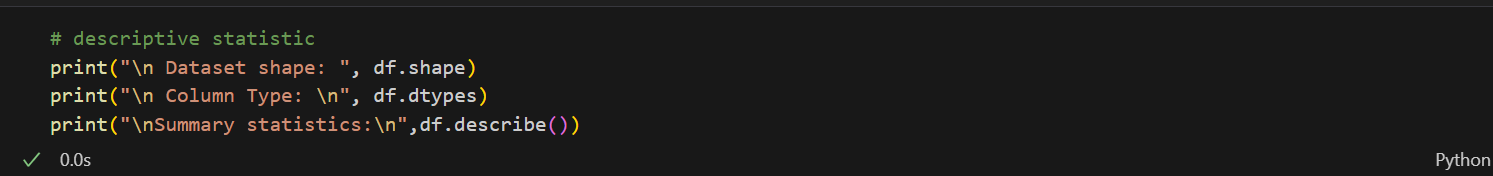
# Data Processing

Our data processing was designed to clean, transform, and prepare the raw data for machine learning algorithms. This section outlines the steps taken to ensure data quality, consistency, and relevance for our analysis.

First, we loaded the dataset (“German Air Fares.csv”) by using Pandas. Pandas provides efficient data structures and data manipulation tools, ideal for handling structured data like our CSV file.



Before going to clean the data, we conducted an exploratory data analysis (EDA) to understand the dataset structure and have an overview of the dataset:



A black rectangular object with white text

Description automatically generated

A screen shot of a computer

Description automatically generated