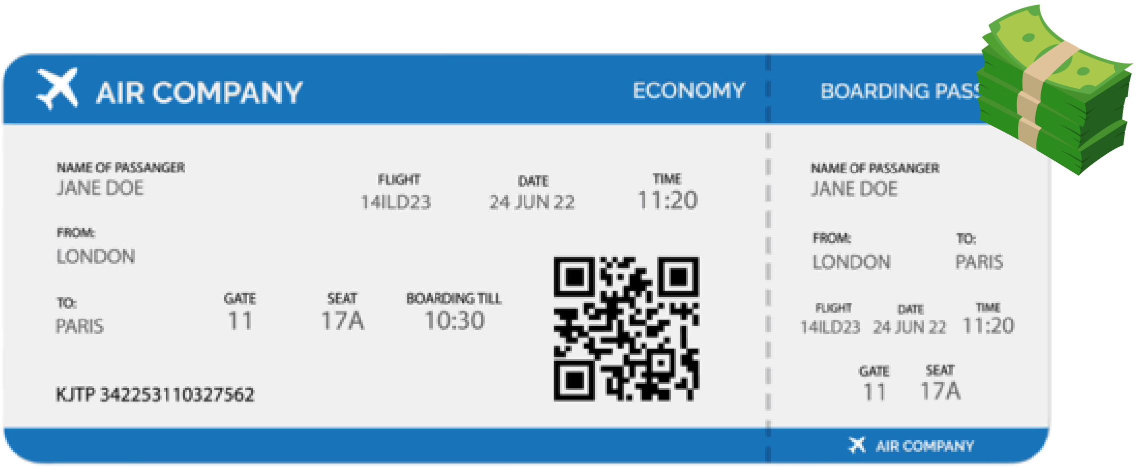
PROJECT PROPOSAL



Predicting the cheapest day to buy a flight ticket using Machine Learning

Boyan Apostolov

20.02.2025

TABLE OF CONTENTS:

1. Introduction

This proposal outlines the research and development of an **AI model** that **predicts** the **optimal time to purchase airline tickets at the <u>lowest</u> price**. The goal is to create a data-driven approach that helps travelers, optimize their booking strategies.
The model will be developed using machine learning techniques and trained on datasets containing ticket **prices**, departure **dates**, and other relevant factors.

Key **Stakeholders**:

- Kalina Bacheva - fellow **student** with a passion of finding **cheap tickets**
- Tanya Apostolova - a **mother** who wants to buy **cheap tickets** for her children

The development of this project will start with data **collection** and data **analysis**, followed by model **training**, **optimization**, and **deployment.**

The **end product** will consist of a page that allows input of a **desired date, departure and arrival airports** and the AI model will output the **predicted date** which will have the **best price option**

---

2. Domain Understanding

2.1 Overview

The airline ticket pricing industry is influenced by **multiple factors**, including:

- **Booking Timeframe** – Prices vary depending on **<u>how far in advance</u>** a ticket is purchased.
- **Departure Date Trends** – Holidays, weekends, and **peak seasons** affect pricing.
- **Route-Specific Trends** – Some routes exhibit more price fluctuations than others.
- **Market Demand** – Airlines adjust prices based on supply and demand.

2.2 Research Question

Which airline pricing practices and buyer behaviors are relevant to predicting the optimal number of days before departure to purchase a ticket?

2.3 Research Methods (DOT Framework)

Exploratory Data Analysis:

- To **analyze** historical **flight pricing data** and **identify trends**.
- To detect **correlations** between booking time, departure dates, and pricing.

2.4 Domain findings

Airline pricing is a **complex field** influenced by **various factors**. Several studies have investigated the relationship between pricing practices, buyer behaviors, and optimal purchase timing:

1. **Dynamic Pricing Strategies**:
    a. *Alderighi et al. (2015)* found that airlines **adjust prices** based on **available seats and time before departure**. Their study of **Ryanair** showed that **prices increase** as **fewer seats remain** and as the **departure date approaches**.
2. **Advance Purchase Discounts (APDs):**
    a. *Gaggero and Piga (2011)* analyzed the UK airline market, confirming the existence of APDs. They observed significant **price increases** at **7**, **14**, and **21** days before departure, suggesting these as **key timing thresholds** for purchases.
3. **Optimal Booking Windows**:
    a. *Etzioni et al. (2003)* developed a data mining approach to **predict airfare changes**. Their research indicated **potential savings** of **27.1%** to **36.9%** by **booking at the right time**, with **optimal windows** varying by route.
4. **Price Volatility and Trends**:
    a. *Williams (2018)* analyzed over **1.5 billion airfare observations**, finding that **prices typically rise substantially** in the **last two weeks before departure**. However, he also noted significant volatility, with an **average of 92 price changes** per **trip**.
5. **Consumer Behavior Impact**:
    a. *Chen and Schwartz (2008)* studied how **consumers' expectations** of **future prices** affect their **booking decisions**. They found that **consumers** often **overestimate** the likelihood of **price decreases**, leading to **suboptimal booking timing.**

These studies collectively suggest that while **general trends exist**, the optimal time to purchase tickets **can vary significantly** based on **specific routes**, airlines, and **market conditions.**

2.5 How my dataset aligns with the findings

Research Insight	What the Research Says	Why My Dataset Is Good
Dynamic Pricing	Prices rise as departure nears and seat availability drops (Alderighi et al., 2015).	My dataset captures **daily prices** leading up to **departure**, reflecting this **dynamic behavior**.
Advance Purchase Discounts (APDs)	**Price jumps** occur at 7, 14, and 21 days before departure (Gaggero & Piga, 2011).	I include "**days before departure**", allowing you to model these **APD thresholds** directly.
Route-specific Booking Windows	**Optimal purchase** timing varies **by route** (Etzioni et al., 2003).	I track departure and arrival airports, enabling **route-level predictions**.
Price Volatility	**Prices** can **change frequently**, especially in the last 2 weeks (Williams, 2018).	**Daily price snapshots** let me capture volatility and fluctuations near departure.
Consumer Booking Behavior	Buyers often misjudge **price trends**, leading to poor timing (Chen & Schwartz, 2008).	My dataset reflects the **real-world view of a buyer**, supporting tools to improve their decision-making.

3. Analytic approach

3. 1 Target variable

The **target variable** in this project is the **number of days before departure when the ticket is the cheapest**.

3.2 Defining success

Success in this project will be defined by:

- **Prediction Accuracy** – The model **correctly identifying** the best booking window for a **significant percentage of cases**..
- **Generalization Across Routes** – The model **should perform well** on **various routes** and travel **periods,** when the necessary data for the route is provided
- **Stakeholder Satisfaction** – **Positive feedback** from the **stakeholders usability** and **effectiveness**

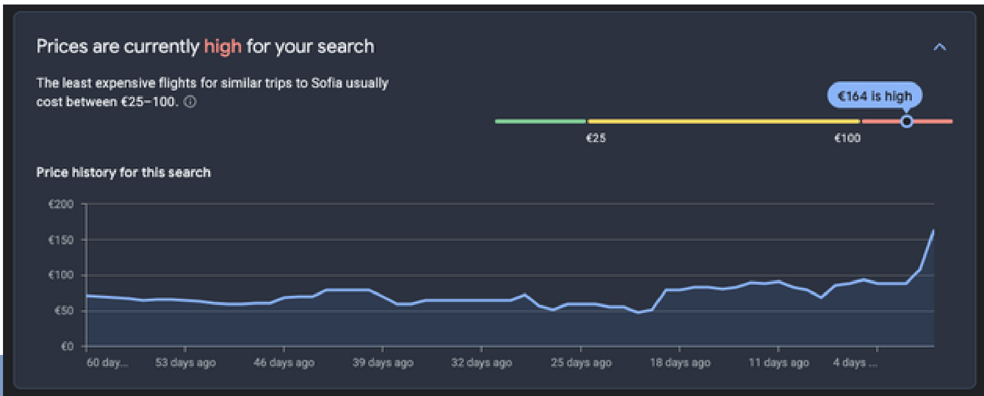
3.3 Feature Selection: Finding Good Indicators

Potentially relevant features include:

- **Days Before Departure** – **How far in advance** the ticket **was booked**.
- **Departure Date** – The **actual date** of the flight.
- Departure **Airport** & Arrival **Airport** – **Route-specific influences**.
- Is the period **Holiday**? – indicating whether the **flight date falls on a holiday.**

3.4 Data Preparation

As I could **not find a suitable dataset for Europe** on the Internet, a **scraper** will be created so I can **gather** the necessary data from **Google Flights,** where we can see the **historical values** for **all flights** for the **last 60 days**.



3.5 Model selection

The current use case requires a regression algorhitm as concrete numerical data is required.

As one of the teachers said, **multiple models** should **be used** and their **results** should **be compared**.

For now, I will **evaluate** and **compare** three machine learning models:

- **KNN** - easies to implement in the **first iterations**
- **Linear Regression** – Selected because it was demonstrated during previous presentations and provides a **simple, interpretable baseline** for predicting the optimal booking window.
- **Decision Trees**

3.6 Model Evaluation Metrics

- **Mean Absolute Error** (MAE) – Measures average prediction error.
- **Root Mean Square Error** (RMSE) – Evaluates model accuracy by penalizing larger errors.
- **R-Squared Score** – Determines how well the model explains price variance.

---

4. Data requirements

4.1 Objectives

- To predict the **best time to book** a flight at the **lowest price**.

4.2 Data requirements

- **Structured Data**: Historical flight booking records containing flight details and prices.

4.3 Data Sources

- **Scraping**: Scraping data historical flight data from **Google Flights**

4.4 Data Legality and Ethics

- Using **only legally available** and **publicly accessible** flight pricing **data**.
- **Avoiding** scraping **personal** or **sensitive customer data**

4.5 Data Diversity

- **Routes**: Ensuring the dataset includes **multiple departure** and arrival airports.
- **Seasonal Variations**: Including flight prices across **different seasons, holidays, and weekdays** vs. **weekends**.
- **Geographical Coverage**: Gathering data from flights **across various regions** for a **well-rounded prediction model**.

4.6 Version Control

Github will be used for dataset/code versioning at the following repositoy: https://github.com/Boyan-Apostolov/Flight-Prices-Predicitons

4.7 Iterative Process

- **Continuously evaluating** the model's **performance** and **refine** data sourcing.
- If model accuracy **is low**, assess and **expand data collection** to **improve coverage**.
- **Monitor price trends** over time, **updating the dataset periodically** to reflect recent market conditions.

5. Planning

Week	**Iteration**	**Details**
Week 2	Iteration 0	Choosing an idea and data gathering
Week 3	Iteration 0	Proposal and iteration zero creation (**K-Nearest-Neighbour**)
Week 4	Iteration 1	Feedback for iteration **zero**, implementing changes.
Week 5 -Week 6	Iteration 1	Implementing **Linear regression**
Week 7	Iteration 2	Feedback for iteration **one**, implementing changes.
Week 8 -Week 9	Iteration 2	Implementing **Decision Trees**
Week 10	Iteration 3	Feedback for iteration **Two**, implementing changes.
Week 11 - Week 12	Iteration 3	Implementing (**third and final algorhitm**)