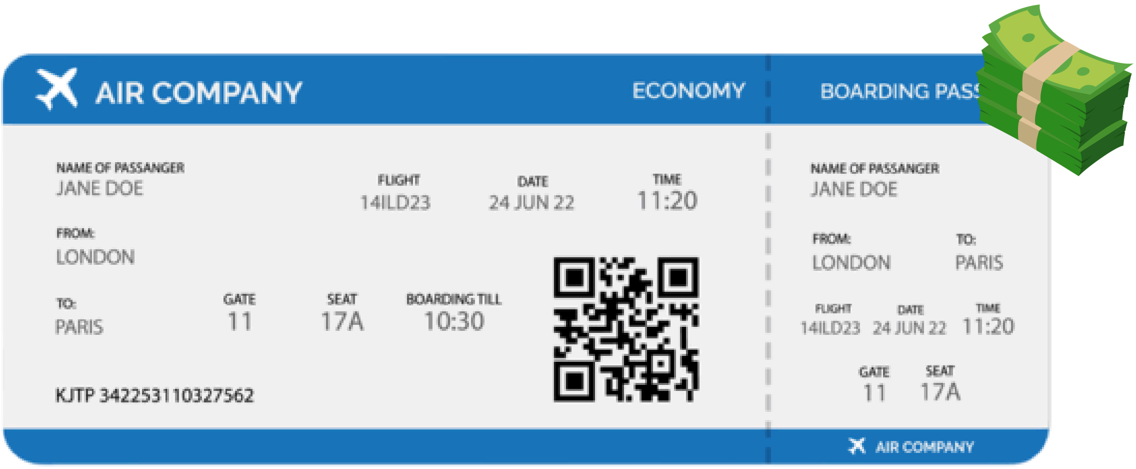
PROJECT PROPOSAL



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 lowest 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 **how far in advance** 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

When is the best time to book a flight to get the cheapest price for a one-way 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.

Model Evaluation:

- To test **different machine learning algorithms** and measure their **accuracy**.
- To **fine-tune** the model and ensure it **generalizes** well.

Prototyping:

- To create an **initial version** of the tool for **testing** (e.g., a simple app or API).
- To **refine usability** based on **feedback**.

2.3 Domain findings

Preliminary research highlights that:

- Flight prices tend to be **lower** when booked **several weeks in advance** but **rise** as **departure nears**.
- **Holidays** and peak seasons significantly **impact** ticket **costs**.
- **Different routes** exhibit distinct pricing patterns.

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**

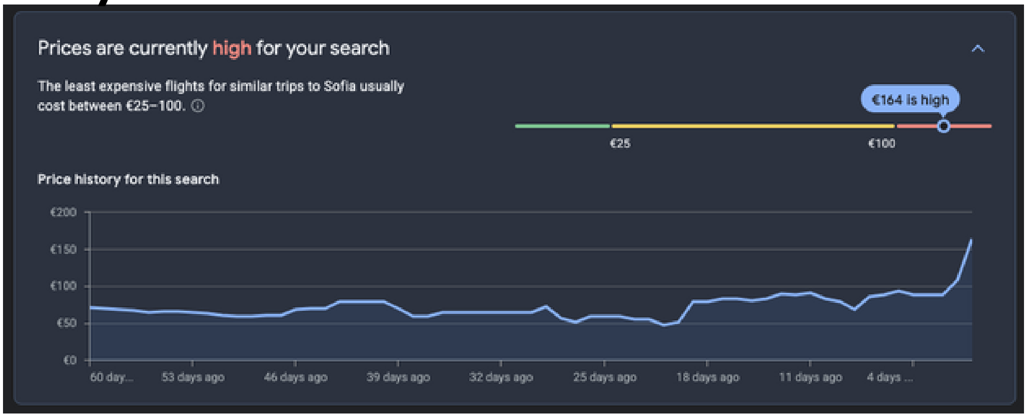
3.3 Feature Selection: Finding Good Indicators

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**.



Prices are currently high for your search

The least expensive flights for similar trips to Sofia usually cost between €25–100.

Price history for this search

Then, the categorical values (e.g. departure and arrival airports) will be converted into numerical values for easier work with the ML algorhitms.

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 7	Iteration 1	Implementing **Linear regression**
Week 8	Iteration 2	Feedback for iteration **one**, implementing changes.
Week 9 - Week 12	Iteration 2	Implementing **Decision Trees**
Week 13	Iteration 3	Feedback for iteration **Two**, implementing changes.
Week 14 - Week 16	Iteration 3	Implementing (**third and final algorhitm**)