Flights delays prediction: Project proposal

Problem Statement

How can travellers predict the likelihood and duration of flight delays with a specified level of accuracy, to build itineraries with connected flights (by the same or different airlines) that meet specific arrival time requirements?

Context

People create flight itineraries at different times, often involving no connections or multiple connections across various airports and flights operated by one or more airlines. Airlines and airports have varying statistics for cancellations and delays, which differ by time of day, season, weather conditions, and high or low traffic periods. Additionally, travellers may have either strict or flexible arrival time requirements, and they may want to evaluate the likelihood of arriving at a specific time or the potential delay duration with a specified level of confidence.

Criteria for Success

A prediction and evaluation model will be developed to provide users with either an estimated arrival time at a specified confidence level or the likelihood of arriving at a specific time for their itinerary.

Scope of Solution Space

The model will be developed exclusively for domestic flights and airports within the United States, based on the available data.

Constraints within the solution space

An itinerary may include up to two connected flights. Each segment of the itinerary must include the date, time (hour), origin, destination, and airline.

Stakeholders to Provide Key Insight or Expertise

Karthik Ramesh, Lead Data (Solution) Engineer

Alexey Kholodov, student

Data sources:

The *US Domestic Flights Delay (2013-2018) dataset*, which includes scheduled and actual departure and arrival times, as reported by certified US air carriers accounting for at least 1 percent of domestic scheduled passenger revenues. The data was collected by the U.S. Office of Airline Information, Bureau of Transportation Statistics (BTS) and covers flights between 2013 and 2018. It includes details such as date, time, origin, destination, airline, distance, and delay status. *(Source: Kaggle)*

Brief Outline of an Approach to the Problem

The project approach will include the following steps:

1. Data analysis, cleaning, and preparation.
2. Exploratory data analysis to observe relationships between various factors in the dataset and evaluate data quality.
3. Identification of potential key factors that may influence the model.
4. Data preprocessing and training set development, including imputing missing values, transforming, encoding, scaling, normalizing data, and splitting into training and testing sets.
5. Application of several models, including tuning hyperparameters for optimal performance.
6. Testing and evaluating the models' results, followed by selecting the final model.
7. Documentation and presentation of results.

These steps outline the initial approach and may be adjusted as the project progresses.

Deliverables

The project deliverables will include a GitHub repository containing:

1. Project report.
2. Slide deck.
3. The US Domestic Flights Delay dataset (2013-2018).
4. Detailed analysis and code for each step of the project, in the form of Python scripts and/or Jupyter Notebooks.