Flight Delay Prediction

1. Problem Statement

Flight delays are a persistent challenge in the aviation industry, costing billions of dollars annually and causing significant frustration for passengers. While some delays are caused by unpredictable events like severe weather, many are driven by systemic factors such as airport congestion, scheduling inefficiencies, and airline operational practices.

The core problem this project addresses is the **uncertainty surrounding flight reliability**. Passengers and stakeholders often lack accessible tools to assess the risk of a specific flight being delayed based on historical patterns and to understand the drivers behind these delays (e.g., is it the airline, the time of day, or the season?).

This project aims to bridge this gap by developing a visual analytics platform that combines interactive storytelling with machine learning to predict flight delay risks and explain the underlying factors.

2. Dataset Overview

Source & Scope: The analysis uses a dataset of US domestic flights from January 2018 to July 2022. Given the large size of the files (over 1 million records each), to ensure computational efficiency while maintaining statistical validity we utilized reservoir sampling to curate a representative dataset of approximately 500,000 flights (100,000 per year) from the raw multi-year data. Despite not being the full dataset, sampling this way allows us to extract samples that are still representative of the overall distribution and that are not biased towards a time-period, airline, etc.

Key Features:

- Temporal: Flight Date, Scheduled Departure Time (CRSDepTime), Month, Day of Week.
- Categorical: Airline (Reporting_Airline), Origin Airport, Destination Airport.
- Operational: Distance, Air Time, Taxi Times (used for analysis, excluded from prediction to prevent leakage).
- Target Variable: IsArrDelayed (Binary: 1 if Arrival Delay > 15 minutes, 0 otherwise).

Basic Exploratory Analysis (EDA):

- Class Imbalance: The dataset is highly imbalanced, with ~83% of flights arriving on time and ~17% delayed.
- The “Long Tail”: While most delays are short, a small percentage of flights experience extreme delays (3+ hours), creating a disproportionate impact on passenger perception.
- Temporal Trends: Delay rates increase steadily throughout the day, peaking in the evening due to cascading effects.
- Seasonality: Distinct peaks in delay/cancellation rates occur in summer (convective weather/volume) and winter (snowstorms), with September-November being the most reliable period according to our data.

3. Business Questions and Objectives

The project was designed to answer the following key questions:

1. What are the primary drivers of flight delays? Does the airline matter more than the route? How does the time of day influence reliability?

2. Can we predict delays using *only* information available before departure? We aim to build a model that does not rely on post-departure features (like Departure Delay or Taxi Out time), which are of course excellent delay predictors, but unavailable at flight reservation time. This way, we're making our application useful for future scheduling, giving it a real use case (plan your flight according to its expected delay).

3. How can we build trust in black-box predictive models? Providing a raw probability is insufficient; users need to know why a flight is flagged as high-risk. We not only want to predict the delay risk, but to discover which patterns drive each individual decision, and why a certain flight is flagged a certain why, and what could we do to increase the odds for a smaller delay (what-if analysis).

<u>Objectives:</u>

- Interactive Exploration: Build a Streamlit dashboard allowing users to filter and visualize data dynamically.
- Predictive Modeling: Develop a robust binary classification model to predict arrival delays (>15 min).
- Explainable AI (XAI): Implement SHAP values and surrogate models to make predictions transparent.

4. Methodology

<u>Data Preprocessing:</u>

- Feature Engineering: Derived features like DepTimeOfDay (Morning, Afternoon, etc.), DayOfWeekName, and seasonality indicators.
- Leakage Prevention: Strictly removed post-departure features (e.g., DepDelay, TaxiOut, ActualElapsedTime) from the training set.
- Handling Imbalance: Used oversampling of the minority class (delayed flights) in the training set to achieve a 50/50 balance, preventing the model from biasing towards the "On-Time" majority. The oversampling was performed by adding delayed samples from the original files (among those that were not sampled originally). The test set remained representative of the real-world distribution. Note that the balanced set was used only for training, the data exploration was used with our real data.

<u>Machine Learning Strategy</u>

- Model Selection: Evaluated Random Forest and XGBoost. XGBoost was selected for its superior performance (ROC-AUC ~0.72) and efficiency. Although an AUC of 0.72 may seem moderate, we actually consider this a very good result for flight delay prediction. Many important factors such as weather and air traffic control are unpredictable before departure. Because the model uses only pre-departure features, it works with limited information. In this context, an AUC above 0.70 shows that the model captures meaningful patterns and can reliably separate high-risk and low-risk flights.
- Evaluation Metrics: Focused on **ROC-AUC** and **Precision-Recall** trade-offs rather than raw accuracy. In imbalanced datasets (83% on-time vs. 17% delayed), accuracy is misleading (a naive "always predict on-time" model would achieve 83% accuracy while being useless). ROC-AUC measures the model's ability to distinguish between classes across all thresholds, while

Precision-Recall focuses on the minority class (delayed flights) that we care most about predicting correctly.

- Threshold Tuning: Implemented a dual-threshold system:
  - Default (0.5): Balanced approach minimizing both false positives and false negatives.
  - High-Recall (0.4): Prioritizes catching delays (reducing false negatives) at the cost of more false alarms, valuable for risk-averse passengers. Threshold selection was informed by Precision-Recall analysis to identify optimal trade-offs.

<u>Visualization & Application (Streamlit)</u>

The solution is deployed as a multi-page Streamlit app:

1. Explore Data: Interactive histograms, heatmaps (Hour × Weekday), and correlation matrices.
2. Key Insights: A curated storytelling page highlighting major findings (e.g., "The Delay Paradox").
3. Predict Delays: An interface for users to input flight details and receive a risk assessment with a gauge chart. They can also see the details on model performance and design choices that were made for total transparency with users.
4. Explainability: Integration of SHAP to show feature contributions (Waterfall plots), feature importance and a Surrogate Decision Tree to visualize decision logic. Note that the decision tree is a rough estimate since the final model uses XGBoost, but it allows users to see very approximately how the decision process go (e.g. if departure time is later than 13:00 and we are in September, predict on-time).

5. Conclusions

The analysis and modeling yield several critical insights:

1. <u>Time is the Critical Factor</u>: Departure time is the strongest predictor of delays. "Cascading effects" mean that early morning flights are significantly safer than evening flights, regardless of the airline or route.
2. <u>Airline Disparity</u>: There is a consistent performance gap (~15-20%) between top-performing airlines and budget/regional carriers, though airline choice remains secondary to temporal factors.
3. <u>Predictability Limits</u>: While the model effectively identifies high-risk scenarios (AUC > 0.7), the inherent randomness of aviation (weather, air traffic control) puts a ceiling on precision. The model is best used as a risk assessment tool rather than a crystal ball.
4. <u>Value of Explainability</u>: By exposing *why* a prediction was made (e.g., "High risk because it's a Friday evening in July"), the tool transforms from a black box into an actionable decision support system for travelers.

Final Recommendation: For maximum reliability, travelers should adopt a three-tier strategy:

(1) Primary: Book early morning flights (before 10 AM) whenever possible. This single factor provides the largest risk reduction.
(2) Secondary: Prefer mid-week travel (Tuesday-Thursday) to avoid weekend congestion.
(3) Tertiary: When flexible, avoid peak operational windows (summer months and winter holidays); September-November offers the most reliable period. This evidence-based approach significantly improves the probability of on-time arrival.

AI Disclosure: This report was drafted with AI assistance to enhance writing quality and structure.