Project Progress Report: Flight Delay Analysis and Prediction Using 2024 US Flight Data

Pratik Narendra Gupta
Western University
pgupta85@uwo.ca

Luke Blommesteyn
Western University
lblommes@uwo.ca

Augusts Zilakovs
Western University
azilakov@uwo.ca

Justin Zomer
Western University
jzomer@uwo.ca

***Abstract*—This report summarizes progress toward our term project in DATASCI 3000A, which analyzes the 2024 US domestic flight dataset to identify patterns in delay behaviour and develop predictive models. This progress update outlines our refined problem definition, measurable objectives, methods implemented to date, and planned next steps.**

I. Problem Definition and Motivation

Flight delays represent a persistent operational and economic challenge within the aviation industry. The US Bureau of Transportation Statistics estimates that flight delays cost billions in lost productivity, increased fuel consumption, and disrupted scheduling across airlines, airports, and passengers. As air travel volume continues to rise, stakeholders increasingly depend on reliable, data-driven tools that can assess operational risk, identify delay patterns, and support better scheduling decisions.

The dataset provided for this project—the 2024 domestic flights dataset from Kaggle—contains detailed information on scheduled and actual departure and arrival times, carrier identifiers, airport metadata, delay durations, and other operational factors. This rich dataset enables a comprehensive investigation into the structural causes and predictors of delays.

Our refined problem statement is: *To identify and characterize the key factors that contribute to departure and arrival delays in US domestic flights during 2024, and to construct a predictive model that reliably estimates whether a flight is likely to be delayed.*

The motivation for this problem goes beyond descriptive analytics: accurate prediction of delay likelihood can support airline operational efficiency, inform customer-facing applications, and contribute to broader optimization efforts across the aviation ecosystem. By analyzing temporal, spatial, and carrier-specific patterns, we aim to produce insights that are both statistically grounded and operationally meaningful.

II. Objectives and Goals

To ensure the project remains targeted and measurable, we structured our work around four well-defined objectives:

1) **Data Cleaning and Preparation:** Develop a fully reproducible cleaning pipeline that addresses missing data, inconsistent time formats, and outliers. Establish a unified dataset with validated delay labels. Completion target: Week 8.
2) **Exploratory Data Analysis (EDA):** Produce a set of visualizations and statistical summaries that characterize delay distributions across carriers, routes, seasons, and times of day. The goal is to generate at least 8–10 high-quality visualizations identifying actionable patterns.
3) **Feature Engineering and Modeling:** Construct predictive models (logistic regression, decision tree, random forest, and at least one advanced method such as XGBoost). Evaluate models using accuracy, F1-score, ROC-AUC, and calibration metrics. Goal: achieve a minimum ROC-AUC of 0.70 on hold-out test data.
4) **Interpretation and Reporting:** Summarize findings that identify the most influential predictors of delays, relate them back to industry context, and provide recommendations for stakeholders such as airlines and airport operations teams.

III. Proposed Methods and Feasibility

Our methodological plan is structured around three progressive stages: preprocessing, exploratory analysis, and predictive modeling. All methods selected have strong feasibility given the dataset size, computational resources available, and course expectations.

A. Data Preprocessing

We implemented an initial cleaning workflow in Python using `pandas`. Key steps completed:

- converted scheduled/actual times into standardized timestamp formats;
- derived delay durations and binary delay labels following common aviation standards (e.g., 15-minute threshold);
- removed impossible or corrupted entries (e.g., negative delays from logging errors);
- identified missing fields and created an imputation plan based on carrier and route-specific historical medians.

These steps ensure a consistent dataset for downstream modeling.

B. Exploratory Data Analysis

We generated early-stage EDA outputs including:

- distributions of departure and arrival delays;
- delay variation across hours of the day and days of the week;
- comparison of delay profiles across major carriers;

- maps illustrating route-level delay patterns.

These visualizations have already revealed strong temporal patterns (e.g., evening delay buildup) that will inform the modeling feature set.

C. Predictive Modeling Plan

For prediction, we will evaluate multiple model families:

- **Baseline:** Logistic regression with regularization to establish interpretability and a simple benchmark.
- **Tree-Based Methods:** Decision tree and random forest to capture nonlinear interactions.
- **Gradient Boosting (XGBoost or LightGBM):** A high-performance model expected to achieve the strongest predictive accuracy.

Model evaluation will use a stratified train-test split and metrics including ROC-AUC, precision, recall, and F1-score. Feature importance, SHAP values, and partial dependency plots will be used to interpret model behaviour.

This methodological plan is fully feasible within the project timeline and provides both rigorous statistical grounding and industry relevance.

IV. Next Steps

Our remaining tasks include:

- finalize imputation strategies and feature engineering (e.g., weather features, airport congestion features);
- expand EDA to include seasonality, holiday effects, and airport-pair clustering;
- train and evaluate all planned models;
- prepare final visualizations and interpretability outputs for the final report.

V. Conclusion

Substantial progress has been made in dataset preparation and early exploratory work. The updated problem definition, measurable objectives, and detailed methodological plan position us well for completing the modeling phase and producing a high-quality final report. Our next steps focus on refining the feature set, training predictive models, and synthesizing insights that connect statistical findings with real operational implications in the aviation industry.