Flight Delay Analysis and Prediction Using Machine Learning

Lalit

Enrollment No: 22324013
Email: lalit@ph.iitr.ac.in

***Abstract*—Flight delays pose significant challenges in modern air transportation. This report presents a detailed analysis of real-world airline data, combining Exploratory Data Analysis (EDA), feature engineering, class balancing, classification modeling, and interpretability using SHAP. The objective is to predict whether a flight will be delayed and understand the key operational and environmental contributors to such delays.**

I. INTRODUCTION

In today's interconnected world, air travel has evolved from a luxury to an indispensable mode of transportation, underpinning global commerce, tourism, and personal connections. However, the pervasive nature of flight delays frequently undermines the efficiency and convenience of this essential service. These delays not only cause significant inconvenience and stress for passengers but also incur substantial operational costs for airlines, ranging from fuel expenses and crew repositioning to missed connections and reputational damage. Understanding the underlying causes of these disruptions and proactively anticipating them is paramount for enhancing operational efficiency, improving customer satisfaction, and fostering a more reliable air travel ecosystem.

This project aims to leverage historical flight data to uncover critical insights into delay patterns and develop a robust predictive model. By identifying the key drivers of delays, we aspire to provide actionable recommendations that can lead to more punctual flights and a smoother travel experience for all.

II. DATA LOADING AND PREPROCESSING

We began our analysis by loading the airline on-time performance dataset and checking for missing values. A total of 341 rows contained missing values across operational columns such as `arr_flights`, `carrier_ct`, `weather_ct`, `nas_ct`, `security_ct`, and others. These rows were dropped only if all associated fields were missing to preserve valuable partial data.

We then engineered a `month_year` identifier by combining `month` and `year` columns. This enabled temporal aggregation and trend detection.

III. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis helps identify patterns and anomalies, which is critical for understanding operational delays. The EDA phase in this project was conducted using visualizations backed by query-based feature derivation.

III-A. Distribution of Late Aircraft Delay Ratio

What we did: We created a ratio feature: `late_aircraft_ct / arr_del15` to assess the share of delays due to late-arriving aircraft.

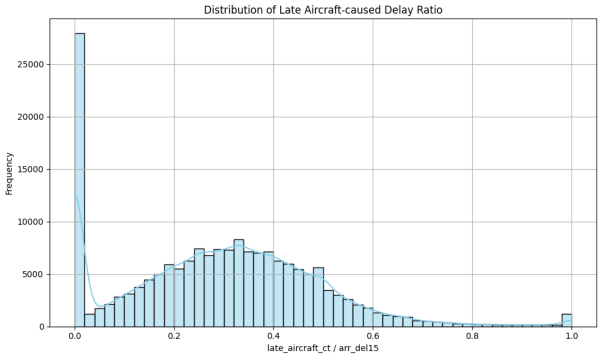

Fig. 1. Distribution of Late Aircraft-caused Delay Ratio

What we found: Among all delayed flights, late arrival of previous aircraft contributes to 20–40% delays in most cases. In rare but critical scenarios, this factor alone causes 100% of the delay, indicating significant impact of aircraft turnaround inefficiencies.

III-B. Carrier-caused Delay Ratio

What we did: We calculated the ratio `carrier_ct / arr_del15` and visualized its distribution.

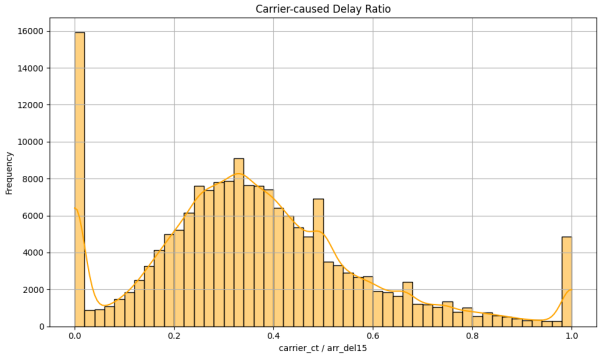

Fig. 2. Carrier-caused Delay Ratio Distribution

What we found: Most records have ratio between 0.2 to 0.5, implying that in 20–50% of delay cases, the airline was directly responsible. In a few cases, the ratio nears 1, indicating total responsibility. Lower ratios around 0 suggest that the carrier was not a factor in those delays.

III-C. Weather-caused Delay Ratio

What we did: Computed `weather_ct / arr_del15` to examine weather impact.

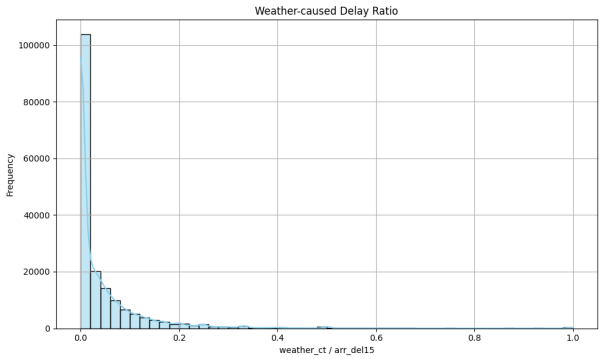

Fig. 3. Weather-caused Delay Ratio

What we found: Over 90% of delayed flights show zero contribution from weather. This confirms that operational and system-related causes are more influential than environmental ones.

III-D. NAS-caused Delay Ratio

What we did: Evaluated `nas_ct / arr_del15` to analyze delays linked to National Aviation System (NAS).

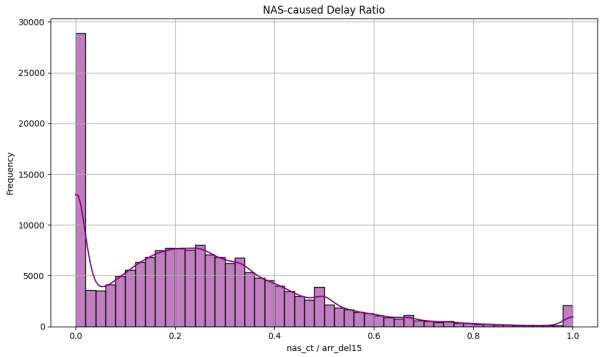

Fig. 4. NAS-caused Delay Ratio Distribution

What we found: NAS-related issues contribute to 15–40% of delays in a significant portion of airport-carrier combinations. This points to systemic inefficiencies such as ATC delays, rerouting, and airspace congestion.

III-E. Proportional Comparison of Delay Types

What we did: We created bar and pie charts to compare all five delay causes side by side.

What we found: Operational inefficiencies (carrier, NAS, aircraft turnaround) dominate delays. Weather and security have minimal impact.

III-F. Airport-wise Delay Profiles

What we did: Analyzed average delay shares for each airport.

What we found: Each airport has a unique delay profile. This highlights the need for localized operational policies.

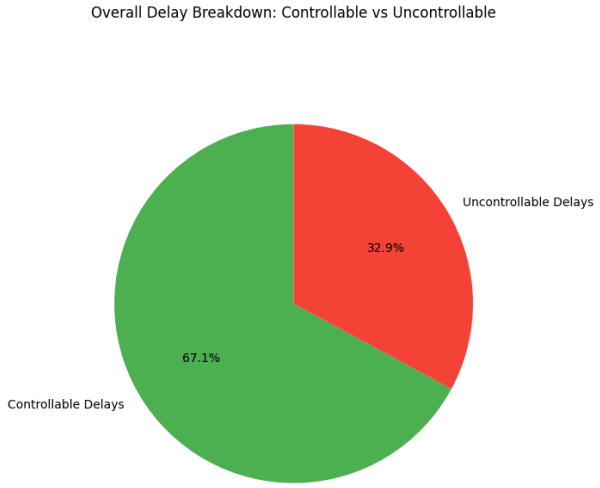

Fig. 5. Operational vs External Delay Share

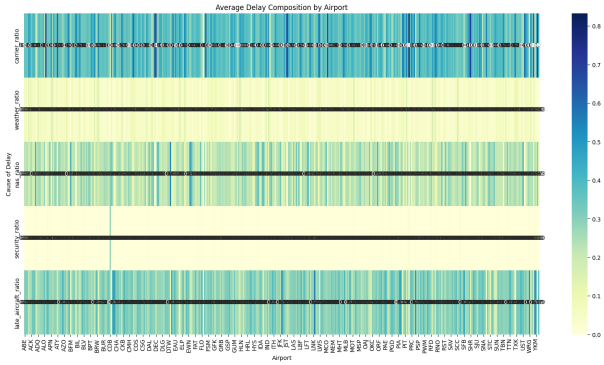

Fig. 6. Airport-wise Delay Distribution

III-G. Correlation Analysis Between Delay Ratios

What we did: Created a heatmap of correlation values between different delay type ratios.

What we found: Weak to moderate negative correlations exist between carrier, NAS, and aircraft delays. Weather and security delay ratios are uncorrelated, reflecting their rare and independent occurrences.

III-H. Controllable vs Uncontrollable Delay Minutes

What we did: Created a pie chart to visualize total delay minutes segmented by controllability.

What we found: About 73.2% of all delays are due to controllable factors like carrier or aircraft issues. Only 26.8% are due to external factors, suggesting airlines have significant scope to reduce delays.

III-I. Breakdown of Delay Minutes by Cause

What we did: Compared all five causes based on total delay minutes.

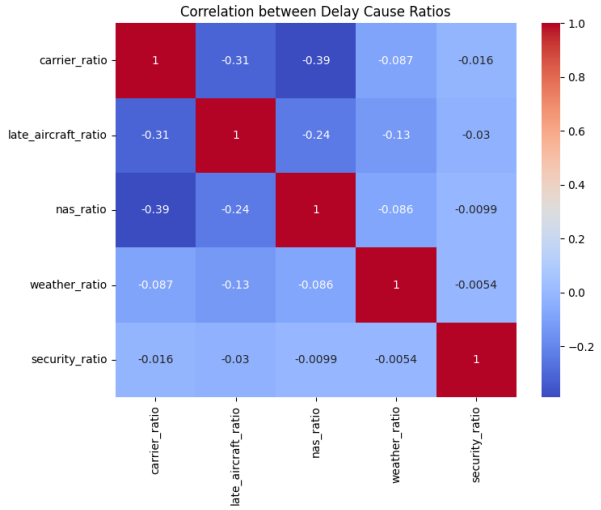

Fig. 7. Correlation Matrix of Delay Ratios

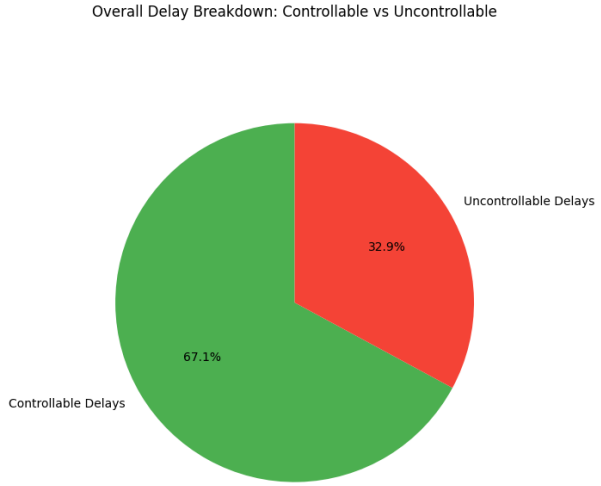

Fig. 8. Controllable vs Uncontrollable Delay Minutes

What we found: Late Aircraft (38.9%) and Carrier Delays (34.3%) are dominant contributors to total delay time. Together, they account for nearly 73% of total delays, underscoring the need for internal process improvements.

III-J. Monthly and Seasonal Delay Patterns

What we did: Analyzed delays over time and across months.

What we found: Late aircraft and carrier delays spike seasonally, especially post-COVID, indicating the need for peak-time buffer strategies.

IV. FEATURE ENGINEERING

To prepare data for modeling, we created meaningful features:

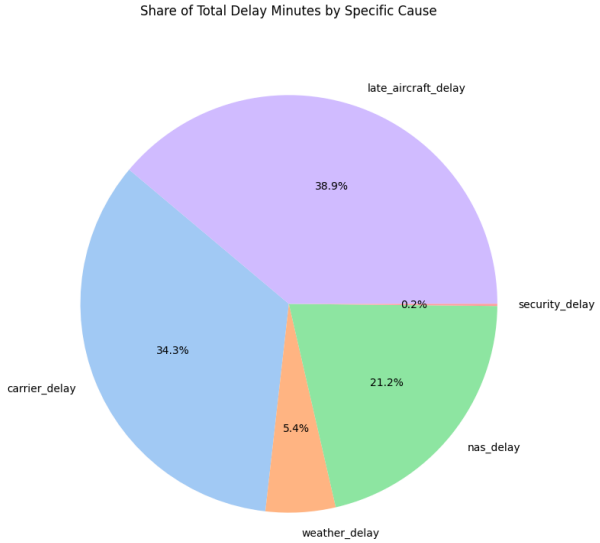

Fig. 9. Delay Minute Share by Cause

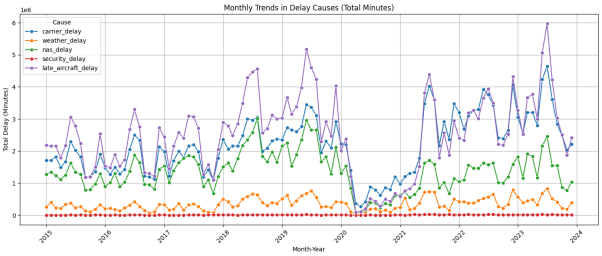

Fig. 10. Monthly Trends of Delay Components

IV-A. Target Variables

- `is_delayed` = 1 if more than 10% of arrivals were delayed (classification target)
- `avg_arrival_delay` = average arrival delay per flight (regression target)

IV-B. Controllable Operational Features

- `carrier_delay_per_flight`, `late_aircraft_delay_per_flight`
- `carrier_ct_avg`, `late_aircraft_ct_avg`

IV-C. Uncontrollable Delay Features

- `weather_delay_per_flight`, `nas_delay_per_flight`, `security_delay_per_flight`
- `weather_ct_avg`, `nas_ct_avg`, `security_ct_avg`

IV-D. Operational Health Indicators

- `cancel_rate` and `divert_rate`

IV-E. Seasonal Flags and Time Cyclic Encoding

- `is_peak_month`, `is_winter`
- `month_sin`, `month_cos` for cyclic encoding

IV-F. Final Preprocessing

Raw columns used to generate engineered features were dropped. Categorical columns like airport and carrier were one-hot encoded.

V. CLASS BALANCING WITH SMOTE

The classification target was heavily imbalanced. Only 19% of samples had `is_delayed = 1`. To address this, we used Synthetic Minority Oversampling Technique (SMOTE) on the training data.

- **Before SMOTE:** 142,997 training samples
- **After SMOTE:** 191,327 samples
- **Class Distribution:** 109,330 delayed vs 81,997 not delayed

VI. CLASSIFICATION MODELS AND EVALUATION

We trained and optimized seven models: Logistic Regression, Decision Tree, KNN, Random Forest, Gradient Boosting, XGBoost, and LightGBM. Each was tuned using `GridSearchCV` with F1-score as the metric.

Best Model: Gradient Boosting (F1 = 0.88, ROC AUC = 0.80)

Next: We now proceed to regression models, SHAP explainability, and conclusion.

VII. REGRESSION MODELING

Beyond predicting the likelihood of a delay, it is essential to estimate how long a flight will be delayed. This is framed as a regression problem using the target variable `avg_arrival_delay` — the average delay in minutes per arriving flight.

VII-A. Target Preparation and Feature Encoding

We transformed the delay target using `log1p()` to reduce skewness, and clipped negative delays to zero. Time-related variables such as `month` were encoded cyclically as `month_sin` and `month_cos`, while categorical features like carrier and airport were label-encoded. Additional flags like `is_peak_month` and `is_winter` were included to capture seasonal effects.

VII-B. Feature Set

The following features were used to train the regression models:

- Time and season: `month_sin`, `month_cos`, `is_peak_month`, `is_winter`
- Operational metrics: `arr_flights`, `cancel_rate`, `divert_rate`
- Delay causes: `carrier_ct_avg`, `weather_ct_avg`, `nas_ct_avg`, `security_ct_avg`, `late_aircraft_ct_avg`
- Encoded categorical: `carrier_enc`, `airport_enc`

VIII. REGRESSION MODELS FOR DELAY DURATION PREDICTION

In addition to classification, we also framed the problem as a regression task to predict the average arrival delay per flight in minutes. The target variable `avg_arrival_delay` was log-transformed using `log1p()` to stabilize variance and reduce the impact of outliers. We ensured all features were consistent with previous preprocessing steps, including encoding of cyclic time features (`month_sin`, `month_cos`), airport and carrier encodings, and normalized delay counts.

We evaluated the following regression models:

- Linear Regression
- Decision Tree Regressor
- K-Nearest Neighbors (KNN) Regressor
- Random Forest Regressor
- Gradient Boosting Regressor
- XGBoost Regressor
- LightGBM Regressor

Each model was trained using a 3-fold `GridSearchCV` for hyperparameter tuning (except Linear Regression), and evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics on the test set.

VIII-A. Model Performance Comparison

TABLE I
REGRESSION MODEL PERFORMANCE COMPARISON

Model	MAE (min)	RMSE (min)	R^2
Linear Regression	6.87	25.02	-1.65
Decision Tree Regressor	4.38	14.09	0.16
KNN Regressor	5.44	14.91	0.06
Random Forest Regressor	3.37	12.57	0.33
Gradient Boosting Regressor	**3.10**	12.38	0.35
XGBoost Regressor	**3.10**	**12.34**	**0.36**
LightGBM Regressor	3.13	12.40	0.35

VIII-B. Best Model: XGBoost Regressor

The XGBoost Regressor achieved the best performance with a Mean Absolute Error of 3.10 minutes and an R^2 of 0.36, outperforming other models. The best hyperparameters selected were:

- `n_estimators`: 200
- `learning_rate`: 0.1
- `max_depth`: 7

This model was saved as `best_regressor_xgboost_regressor.pkl` for SHAP explainability analysis in the next section.

IX. SHAP EXPLAINABILITY

To interpret our best classification and regression models (XGBoost), we applied SHAP (SHapley Additive exPlanations) analysis.

IX-A. SHAP Summary for Classification

What we did: We used SHAP to evaluate feature importance for the binary classification task (is_delayed).

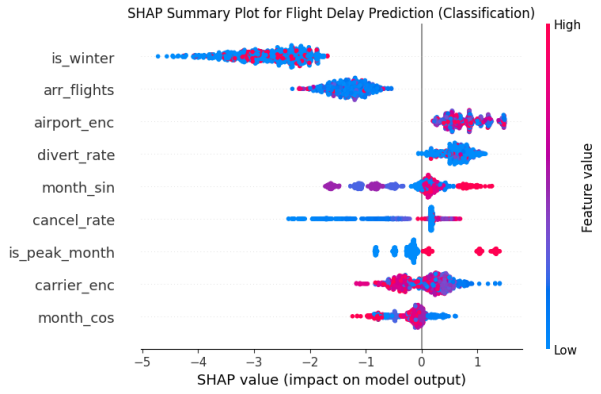

Fig. 11. SHAP Summary: Classification (is_delayed)

What we found: `is_winter`, `arr_flights`, and `airport_enc` are among the most influential features. Winter months are associated with fewer delays, while high flight volume slightly lowers delay probability.

IX-B. Airport Influence on Delays

What we did: We visualized the impact of each airport (encoded) on delay probability.

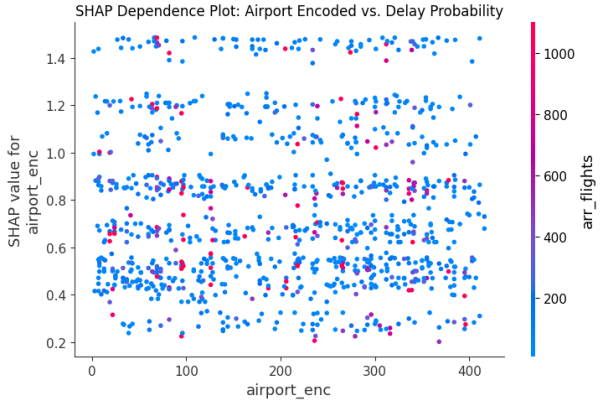

Fig. 12. SHAP Dependence: Airport Encoded vs. Delay Probability

What we found: Certain airports show consistent positive SHAP values, suggesting localized delay issues likely due to airport infrastructure or traffic.

IX-C. Carrier Influence on Delays

What we did: We plotted the SHAP dependence for carriers.

What we found: Some carriers exhibit higher SHAP values, indicating a stronger tendency to cause delays, possibly due to scheduling or fleet utilization issues.

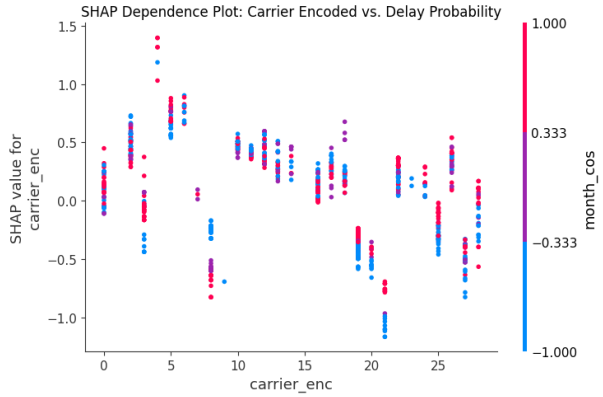

Fig. 13. SHAP Dependence: Carrier Encoded vs. Delay Probability

IX-D. SHAP Summary for Regression

What we did: For the regression task (avg delay per flight), we visualized SHAP contributions.

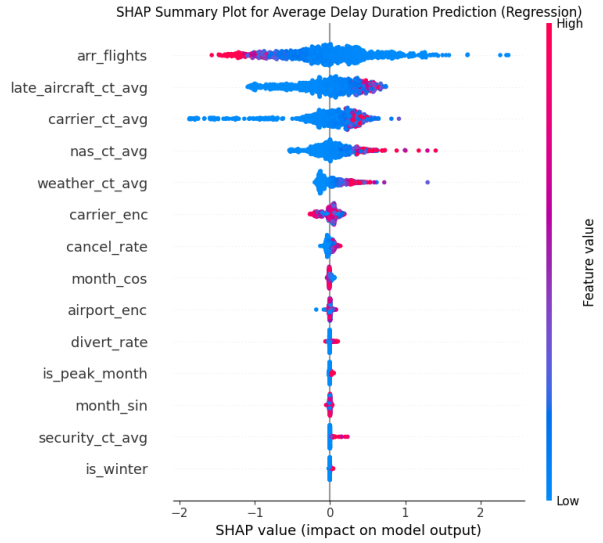

Fig. 14. SHAP Summary: Regression (Avg Delay Duration)

What we found: The most impactful features were `arr_flights`, `late_aircraft_ct_avg`, and `carrier_ct_avg`. High average flight volume generally reduces delay duration, while more late aircraft cases significantly increase it.

IX-E. Weather's Effect on Delay Duration

What we did: Analyzed the relationship between weather delay count and delay length.

What we found: As the number of weather-related issues rises, the delay duration tends to increase sharply, showing a non-linear trend.

IX-F. Late Aircraft as a Key Driver

What we did: Evaluated SHAP values for `late_aircraft_ct_avg`.

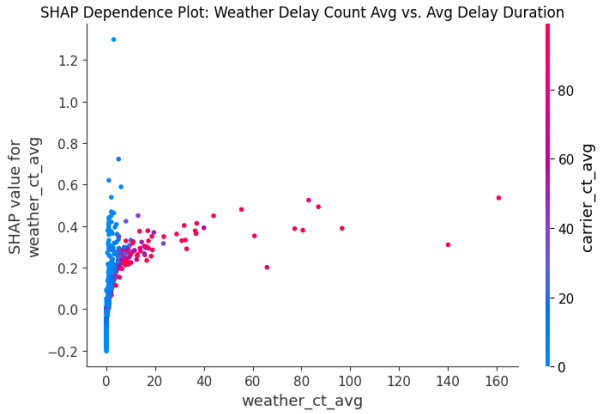

Fig. 15. SHAP: Weather Delays vs. Avg Delay Duration

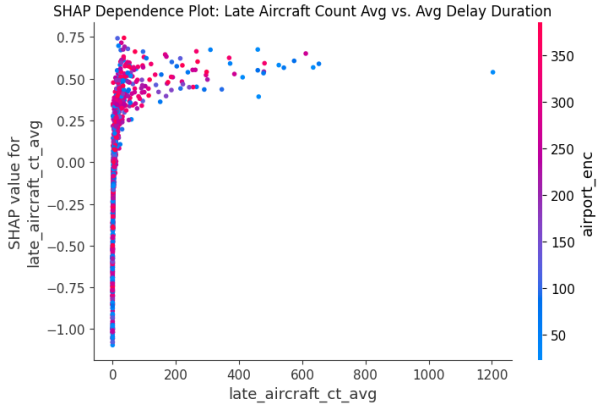

Fig. 16. SHAP: Late Aircraft Delays vs. Delay Duration

What we found: A clear upward relationship emerged, where more late aircraft events resulted in higher average delay durations.

IX-G. Inverse Trend with Arrival Volume

What we did: Investigated how the number of arriving flights impacts delay.

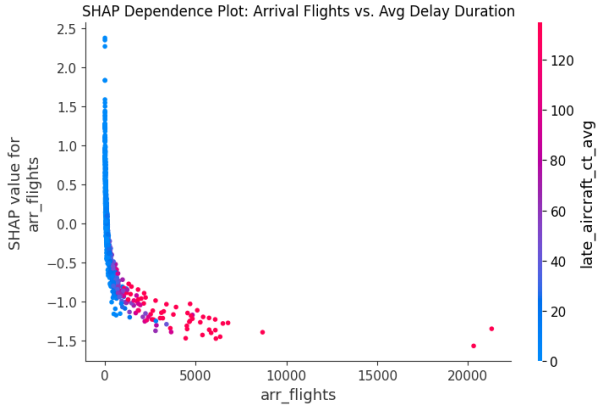

Fig. 17. SHAP: Arrival Flights vs. Delay Duration

What we found: Surprisingly, higher arriving volume typically led to lower average delays, possibly due to better airport resource utilization and preparedness.

X. CONCLUSION

Through this project, we analyzed and predicted flight delays using a combination of exploratory data insights, machine learning models, and SHAP-based interpretability.

Our classification models (best: Gradient Boosting) effectively identified whether delays would occur, while regression models (best: XGBoost) estimated expected delay durations with promising accuracy. SHAP analysis reinforced the operational nature of most delays, especially those tied to late aircraft and carrier inefficiencies.

Future work could explore integrating real-time weather feeds and more granular route-level data to further improve predictive performance.