Flight Delay Analysis

Krrish Jindal , 23323020

Contents

Problem Overview & Method Summary

Objective

Analyze 2015–2023 US airline delay data to understand delay causes and predict delays. We perform two tasks: (1) binary classification of each flight-month as **“Delayed”** (>15 min) or on-time, and (2) regression to predict **delay duration** (minutes) for each flight..

Approach

Clean and merge flight and delay-cause data; create an *Operational Adjustability Index (OAI)* to weight controllable delays higher. We split the data into train/test sets and train XGBoost models for classification and regression, using OAI-based sample weights. Models are evaluated on held-out data (confusion matrix, ROC, MAE/RMSE, R^2).

Exploratory Data Analysis (EDA)

Step 1

Import, clean, and merge datasets from multiple sources (flight-level + delay causes).

Step 2

Handle missing values and engineer new features like delay flags and total delay

Step 3

Aggregate delay statistics across months, years, carriers, and airports.

Step 4

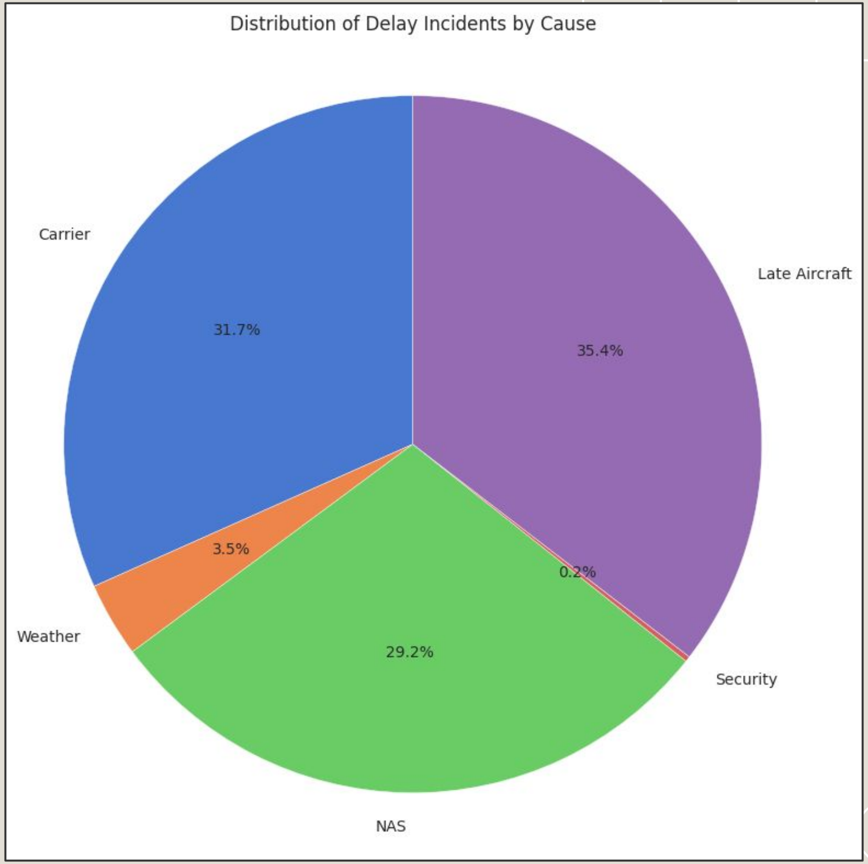
Visualize delay distributions, cause breakdowns, and seasonal delay trends.

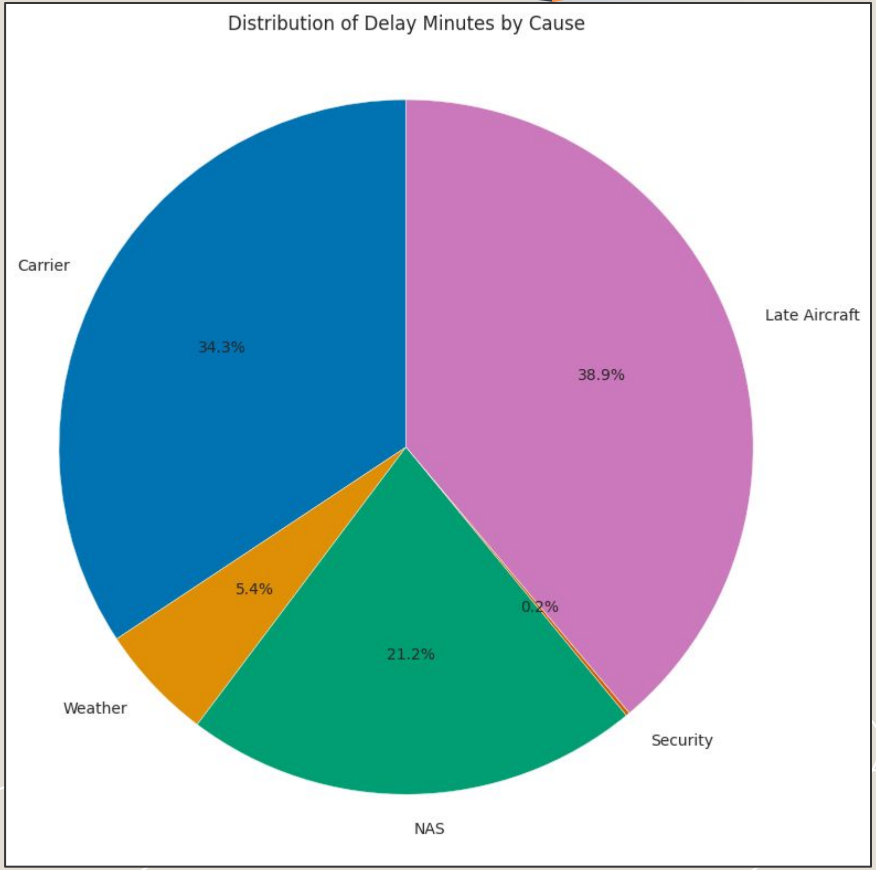
Step 5

Identify top delay contributors (late aircraft, carrier) and uncover seasonal peaks

Goal

Establish a solid understanding of delay patterns to inform modeling and strategy


Distribution of Delay Incidents by Cause
Carrier
31.7%
Late Aircraft
35.4%
3.5%
0.2%
Weather
29.2%
Security
NAS


Distribution of Delay Minutes by Cause
Carrier
34.3%
Late Aircraft
38.9%
5.4%
0.2%
21.2%
Weather
Security
NAS

Temporal Pattern Analysis Summary

Seasonality

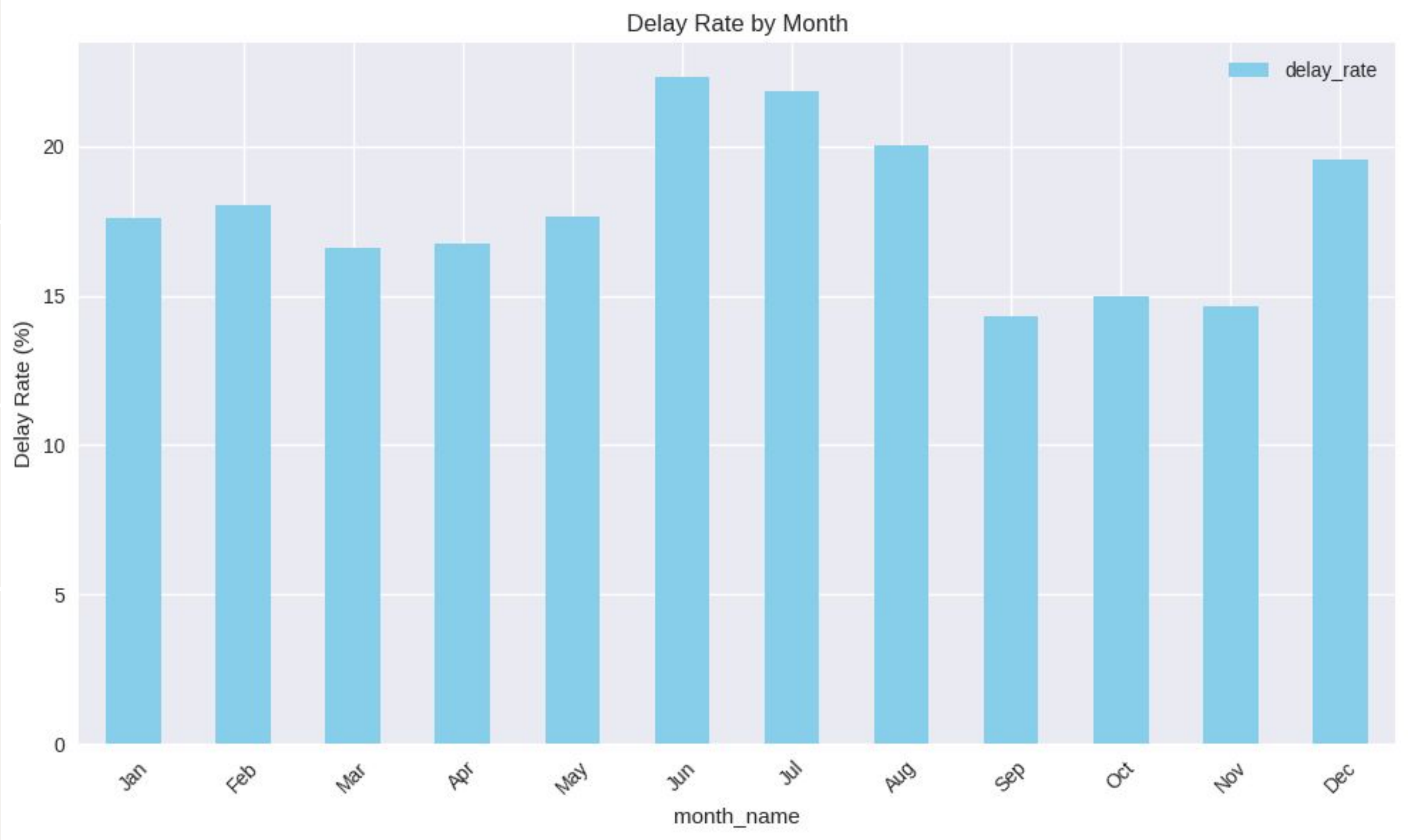
- Delay rates follow a clear seasonal cycle across years.
- Summer (June–August) consistently shows highest delays.
- **Worst month: June (~21.9% delayed flights)**
- **Lowest Delay month: September (~14.5%)**
- **Fall and winter months have better on-time performance.**

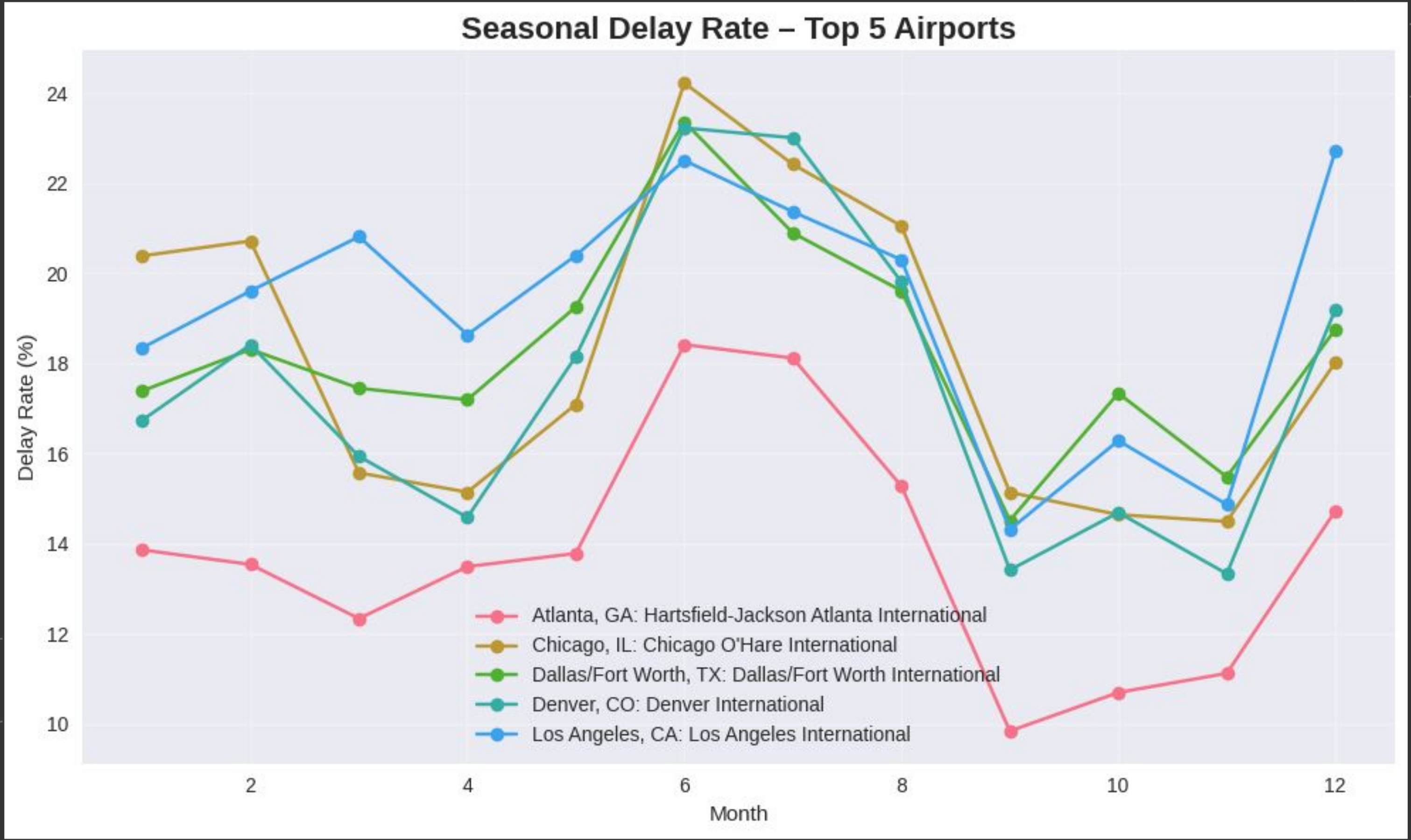
Trend

- From 2015 to 2023, delay rate fell by ~4.9 percentage points.
- Meanwhile, flight volume rose ~29%, indicating efficiency gains.
- **Busiest month: March (5.2 Million Flights)**
- **Seasonal load patterns strongly impact delays — month is a top predictive feature.**

Implication

- Time-based patterns are critical in modeling and planning.
- Reinforces the need to include month/season in ML features.
- These patterns validate seasonality as a major delay driver


Delay Rate by Month
delay_rate
Delay Rate (%)
0
5
10
15
20
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
month_name



Seasonal Delay Rate – Top 5 Airports
Delay Rate (%)
24
22
20
18
16
14
12
10
Month
2
4
6
8
10
12
Atlanta, GA: Hartsfield-Jackson Atlanta International
Chicago, IL: Chicago O'Hare International
Dallas/Fort Worth, TX: Dallas/Fort Worth International
Denver, CO: Denver International
Los Angeles, CA: Los Angeles International

Predictive Modeling & Performance

TRAINING

We trained XGBoost models for both tasks, using OAI-derived weights to emphasize controllable delays. Classification predicts Delayed vs On-time; regression predicts minutes of delay.

01

CLASSIFICATION RESULTS

On test data, accuracy ≈86.9%, precision ≈70.9%, recall ≈85.8%, F1 ≈77.7%, and ROC-AUC ≈0.944 (The high recall indicates most delayed flights are correctly identified.)

02

REGRESSION RESULTS

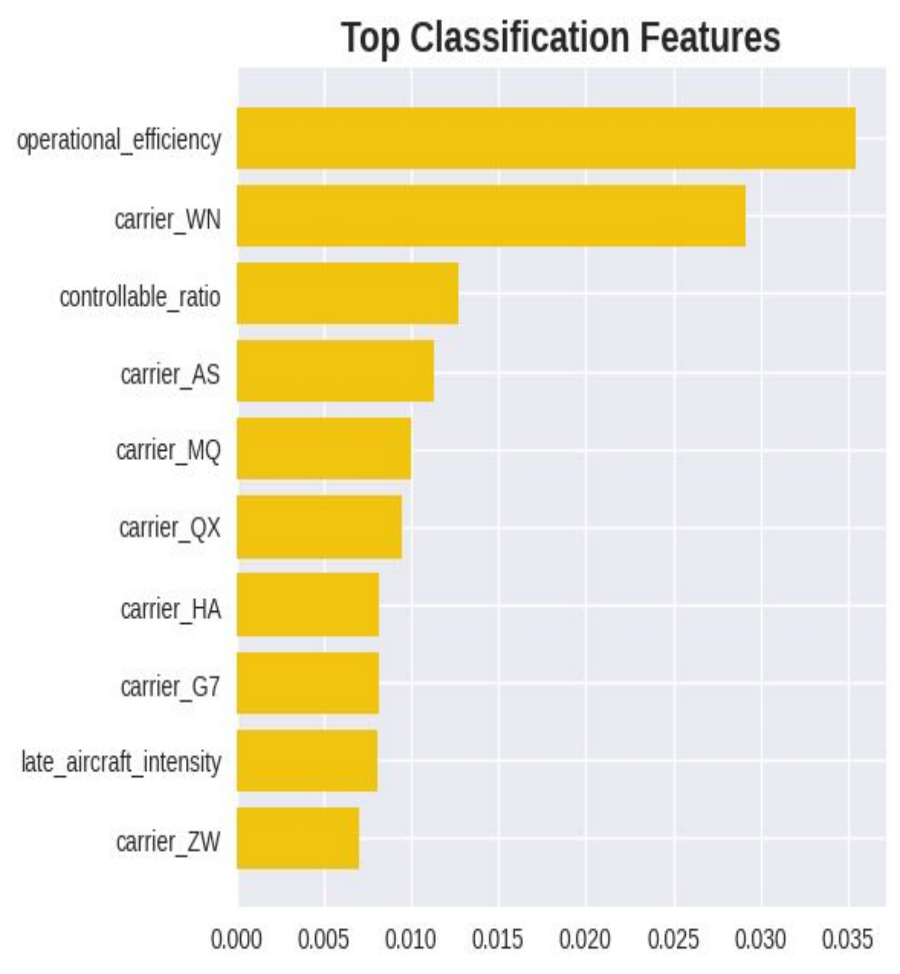
Delay-duration prediction has MAE ≈3.35 min, RMSE ≈12.84 min, R^2 ≈0.320 These errors are moderate given the wide variation in delays. Residual analysis shows most predictions near the truth with some under/over-estimation of extreme delays.

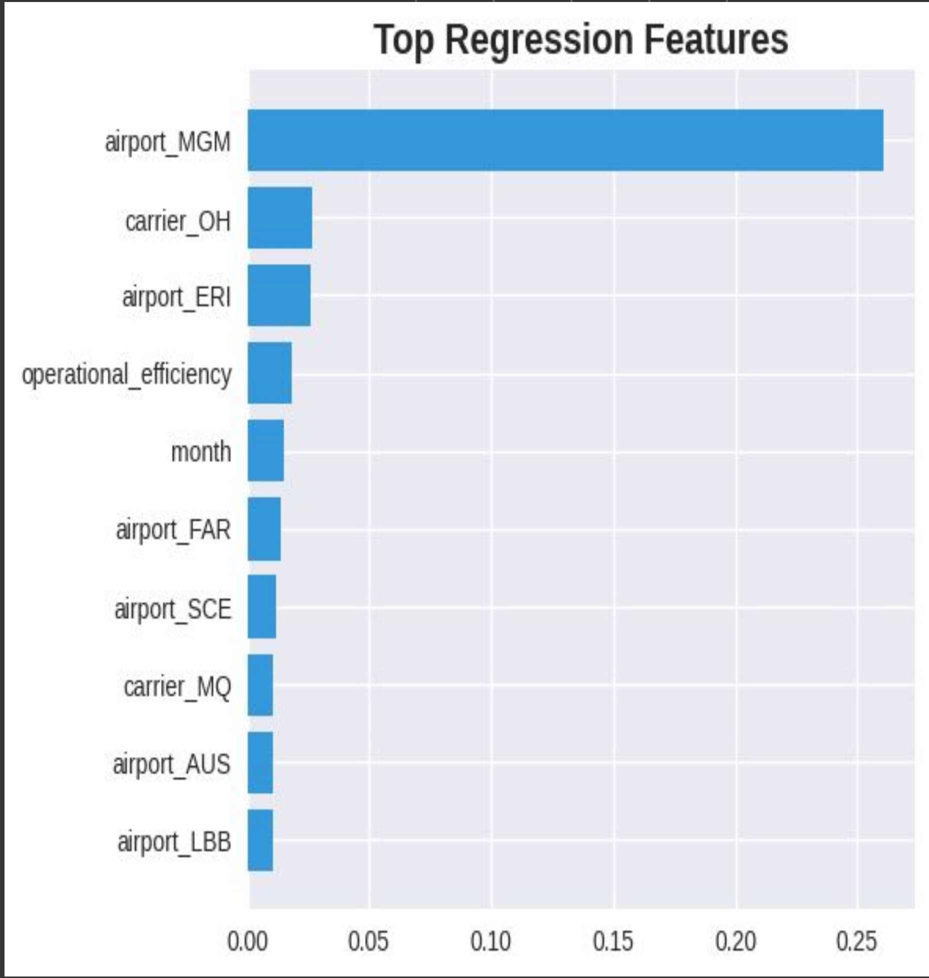
03

VALIDATION

We use a held-out test set and report confusion matrix / ROC for classification. The ROC curve confirms the model greatly outperforms random guessing (AUC≈0.94)

04


Top Classification Features
operational_efficiency
carrier_WN
controllable_ratio
carrier_AS
carrier_MQ
carrier_QX
carrier_HA
carrier_G7
late_aircraft_intensity
carrier_ZW
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035


Top Regression Features
airport_MGM
carrier_OH
airport_ERI
operational_efficiency
month
airport_FAR
airport_SCE
carrier_MQ
airport_AUS
airport_LBB
0.00
0.05
0.10
0.15
0.20
0.25

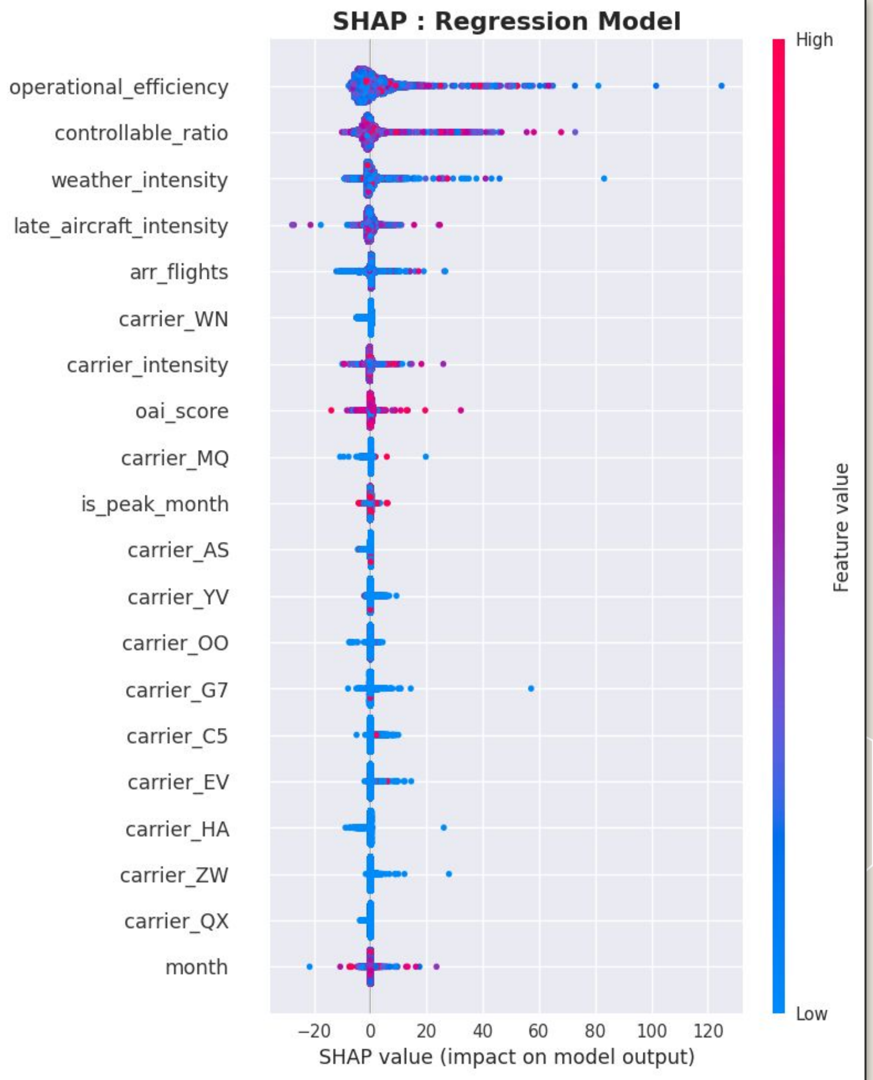
SHAP & Feature Importance

Key drivers of delays:

SHAP analysis of the classifier highlights *seasonal factors* as the strongest predictors. In particular, **Month** and **Year** (capturing seasonality and trends) have the largest SHAP values, meaning they contribute most to predicting delays. Among carrier features, airlines like **Delta (DL)**, **Southwest (WN)** and **Alaska (AS)** emerge as important (higher predicted delay risk).

Other features:

For the regression model, features such as **flight volume (arr_flights)** and **late-aircraft intensity** (delay propagation) are among the top contributors to predicted delay minutes. In general, controllable operational factors rank higher than uncontrollable ones



Operational Adjustability Index (OAI) Summary

Focus on Controllable Delays

1

What is OAI?

OAI is a custom metric that gives more weight to controllable delay types (e.g. carrier delays, late aircraft), helping highlight actionable delay causes.

2

Modeling Impact

With OAI-weighted regression, weighted MAE improved from 3.35 min to 2.65 min, proving better prediction for controllable delays.

3

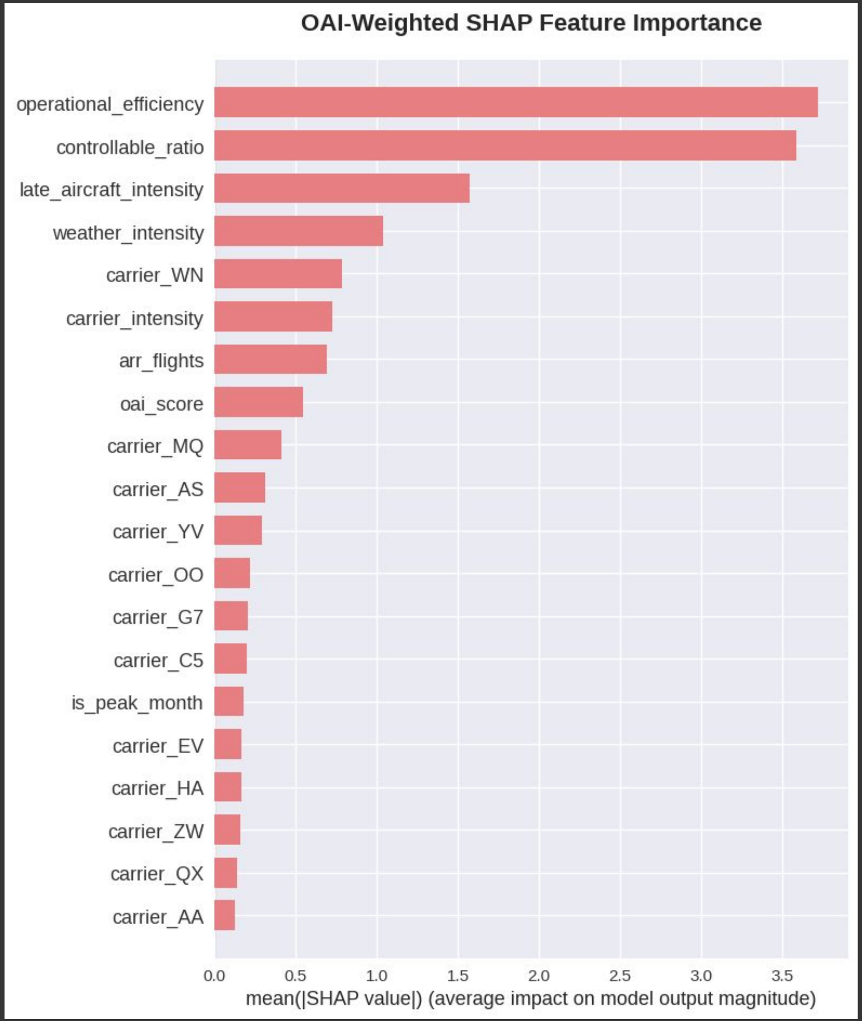
Key Implications

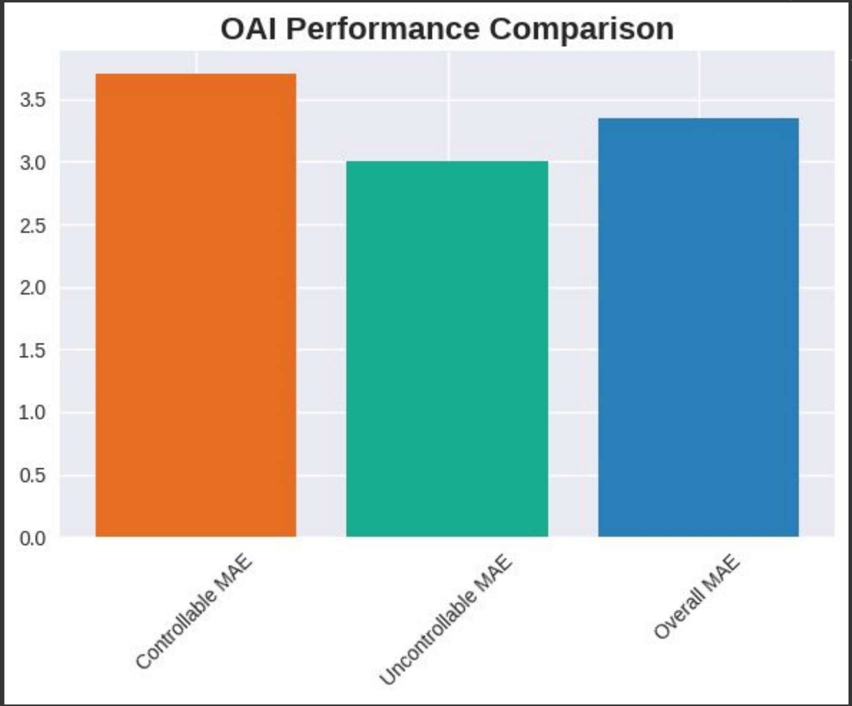
Carrier-related delays gain more importance under OAI, showing a strong need for better crew scheduling, aircraft turnaround, and resource planning.

4

Recommendations

Focus on reducing controllable delays during peak periods (e.g. summer): buffer schedules, rapid maintenance, and optimize gate/staff allocation.


OAI-Weighted SHAP Feature Importance
operational_efficiency
controllable_ratio
late_aircraft_intensity
weather_intensity
carrier_WN
carrier_intensity
arr_flights
oai_score
carrier_MQ
carrier_AS
carrier_YV
carrier_OO
carrier_G7
carrier_C5
is_peak_month
carrier_EV
carrier_HA
carrier_ZW
carrier_QX
carrier_AA
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
mean(|SHAP value|) (average impact on model output magnitude)


OAI Performance Comparison
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Controllable MAE
Uncontrollable MAE
Overall MAE

Thank you