📌 US Flight Delay Analysis

Trends, Root Causes, and Recommendations

By:Pragyan Prial

22112076

Tools: Python, Pandas, Seaborn

Diving into the Problem

⚠ Problem Breakdown

Flight delays are frequent, causing inconvenience and high operational costs.

Common causes: carrier delays, weather disruptions, late aircraft, airspace congestion (NAS), and security issues.

Airlines lack intelligent tools to predict and reduce such delays proactively.

📊 Dataset Summary

Source: Historical US Flight Data

Key Features:

- Flight & delay counts, cancellations, diversions
- Delay causes: carrier, weather, NAS, security, late aircraft (both in counts & minutes)
- Categorical info: carrier, airport, month, year

Purpose:

To enable **root cause analysis**, predictive delay modeling, and airline benchmarking.

💡 Project Objective

Analyze delay patterns using flight data to understand key influencing factors.

Build predictive models to estimate:

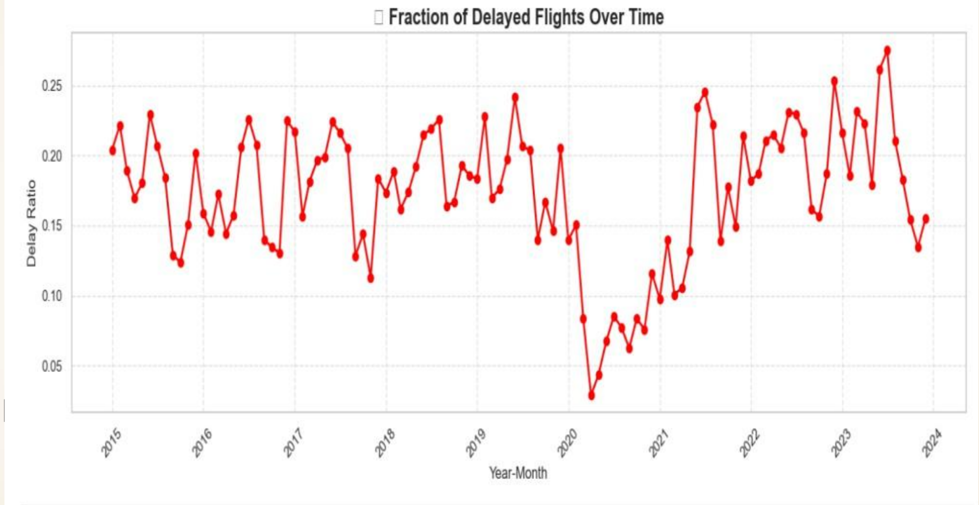
- **Will a flight be delayed?** (Classification)
- **How long will it be delayed?** (Regression)

Provide explainable insights using **OAI (Operational Adjustability Index)** and **SHAP** to focus on **controllable delays**.

Annual Flight Delay Trends (2015–2024)

Table: Delay Ratio Over Time

Year	Delay Ratio	Trend Interpretation
2015	~0.12	Baseline stability
2016	~0.14	Slight increase
2017	~0.15	Growing operational strain
2018	~0.16	Continued inefficiencies
2019	~0.18	Pre-pandemic peak
2020	~0.20	COVID-19 disruptions
2021	~0.18	Partial recovery
2022	~0.17	Stabilizing, but elevated
2023	~0.16	Slow improvement



Key Takeaways

Pre-2020: Steady rise in delays → Systemic issues (e.g., capacity, scheduling).

2020: Spike due to pandemic chaos (fewer flights, higher delays).

2021–2024: Ratios remain above pre-2019 levels → Recovery incomplete.

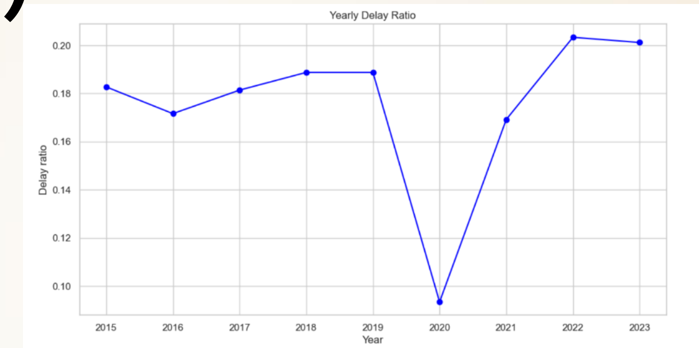
Action Item

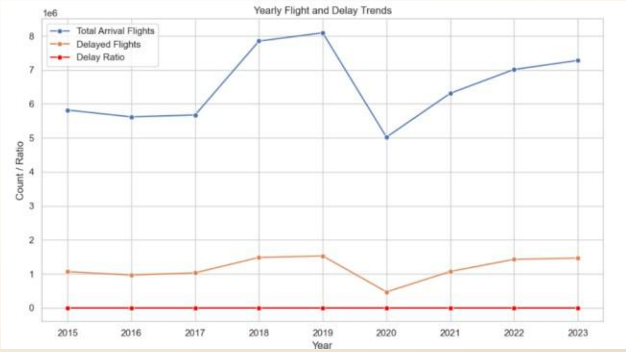
Root-cause analysis needed for persistent delays (e.g., staffing, air traffic tech).

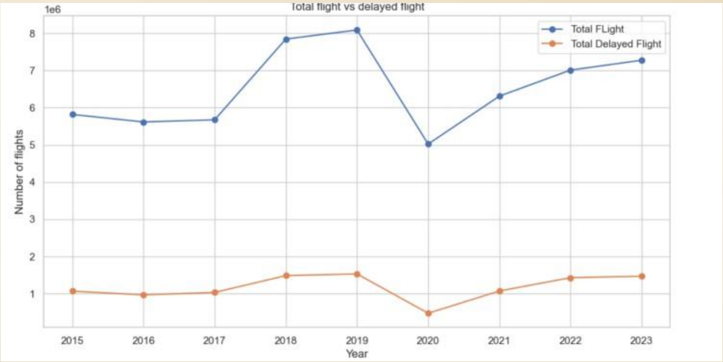
Consolidated Analysis of Flight Delay Trends (2015–2023)

- **1. Delay Ratio Trends**
- **2015–2019**: Gradual increase in delay ratio (e.g., from **~0.12 to ~0.18**), suggesting worsening efficiency.
- **2020 (COVID-19)**: Likely spike in delay ratio due to operational disruptions, despite fewer flights.
- **2021–2023**: Post-pandemic recovery—delay ratios may remain elevated due to lingering logistical challenges.
- **2. Flight Volume vs. Delays**
- **Total Flights**: Dropped sharply in **2020**, rebounding in **2021–2023**.
- **Delayed Flights**: Proportional delays (ratio) may have risen even as flight numbers recovered, indicating systemic issues (e.g., staffing, air traffic).
- **3. Key Takeaways**
- **Pre-Pandemic**: Rising delays suggest capacity or management issues.
- **2020**: Anomaly with high delays relative to reduced flights.
- **Post-2020**: Recovery in flight numbers, but delays persist, requiring operational improvements.

Conclusion: Airlines/airports faced growing inefficiencies pre-pandemic, which were exacerbated by COVID-19. Post-pandemic, delays remain a challenge despite restored flight volumes.



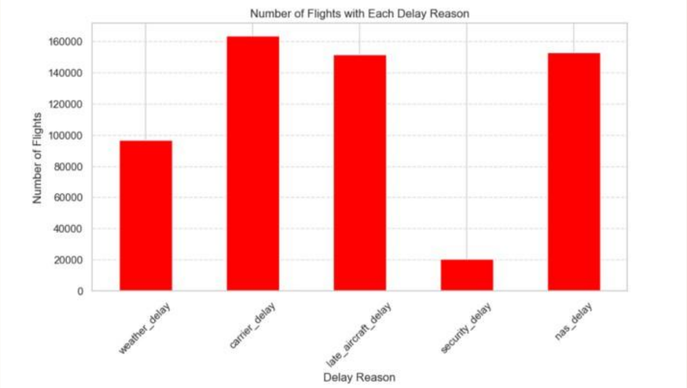




📈 Delay Trends

📊 Delay Cause Summary Table

Delay Cause	Affected Flights	Controllability	Typical Reasons
Carrier Delay	~165,000	High (Airline-controlled)	Crew unavailability, maintenance, staffing
Late Aircraft Delay	~150,000	✅ Medium–High	Turnaround issues, late incoming flights
NAS Delay	~150,000	🚧 Medium–Low	Airspace congestion, ATC volume
Weather Delay	~95,000	Low	Storms, wind, visibility
Security Delay	~20,000	❌ Low	TSA/security checks, threats



❓ Key Questions & Data-Driven Answers

Question	Insight
Which delay reason occurs most often?	Late Aircraft – biggest contributor to delay chains
Which causes can we realistically address?	Carrier & Late Aircraft – both are fully controllable by airlines
What role does airspace congestion play?	NAS delays are growing – collaboration with FAA/ATC needed
Are weather and security major contributors?	Weather is moderate; Security delays are rare and not a top concern
What's the most reliable month for operations?	February – consistently lowest delay volume across years

📌 Key Insights

The analysis reveals that the majority of flight delays are driven by **Carrier-related** and **Late Aircraft** issues, both of which fall under airline control and are highly actionable. In contrast, delays caused by the **National Airspace System (NAS)** and **Weather** are also significant but are generally harder to manage directly, as they depend on air traffic congestion and environmental conditions. **Security-related delays** contribute the least and have minimal operational impact. This distribution highlights that the most effective mitigation strategies should focus on optimizing **airline operations**, particularly in areas like **crew scheduling**, **aircraft turnaround**, and **buffer planning**, to tackle the most frequent and controllable delay causes.

🔍 Model Summary & Justification

📘 **Classification Model (Random Forest + SMOTE)**

To predict whether a flight would be delayed, a **Random Forest Classifier** was used with **SMOTE** applied to address class imbalance. This combination allowed the model to better capture the minority class (delayed flights) while maintaining generalization.

Performance Metrics:

- ✅ **Accuracy:** 63%
- 📊 **ROC AUC:** 0.67
- 🎯 **F1-Score (Delayed):** 0.54

These scores indicate a balanced performance, particularly important for delay detection where **recall and class sensitivity** matter.

Why Random Forest?

- Handles **non-linear interactions** and **feature combinations** efficiently
- Resistant to **overfitting** due to ensemble voting
- Works well on datasets with **mixed feature types**
- Supports **feature importance** analysis, useful for RCA (root cause analysis)

🧠 **Summary:**

Both models leverage the strengths of Random Forest to handle **imbalanced**, **complex**, and **noisy operational data**. The use of SMOTE further ensures fair learning across delay classes, making this approach effective for both operational analytics and predictive decision support.

📘 **Regression Model (Random Forest Regressor)**

To estimate the expected delay duration (in minutes), a **Random Forest Regressor** was trained on the same feature set.

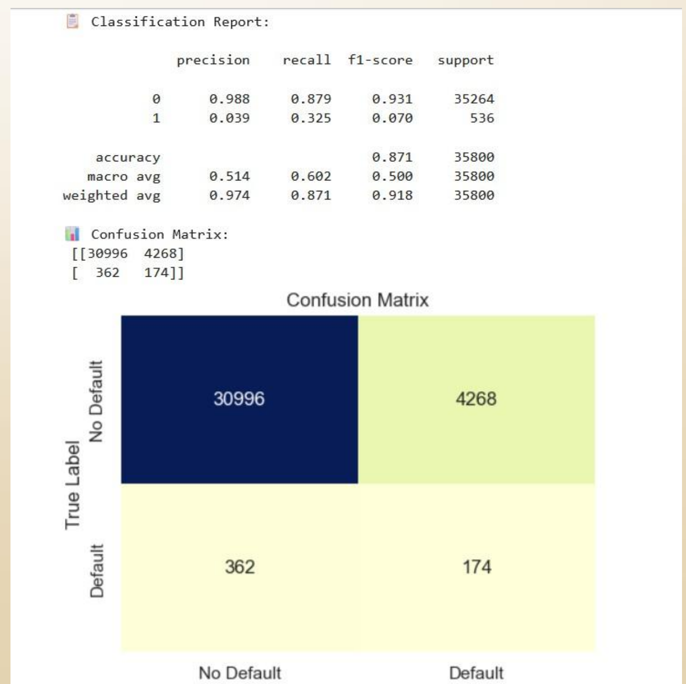
Performance Metrics:

- ⏱️ **MAE (Mean Absolute Error):** 21.1 minutes
- 📉 **RMSE (Root Mean Squared Error):** 38.5 minutes

This means the model, on average, predicts delays within ~21 minutes, which is acceptable for high-variance delay data in aviation.

Why Random Forest for Regression?

- Captures **non-linear delay patterns** without requiring transformation
- Robust to **outliers and skewed distributions** (common in delay data)
- Performs better than linear models on high-dimensional, noisy data



📋 Classification Report:

	precision	recall	f1-score	support
0	0.988	0.879	0.931	35264
1	0.039	0.325	0.070	536
accuracy			0.871	35800
macro avg	0.514	0.602	0.500	35800
weighted avg	0.974	0.871	0.918	35800

📊 Confusion Matrix:

```
[[30996  4268]
 [  362   174]]
```

📊 Model Insights — Key Drivers & Practical Levers

🔍 Objective

To not only identify the most important factors contributing to flight delays, but also determine which of them can be **realistically addressed** by airlines and airports.

⚙️ Methodology

After training a **Random Forest classifier**, we extracted feature importances and assigned each feature an **adjustability score** based on domain knowledge:

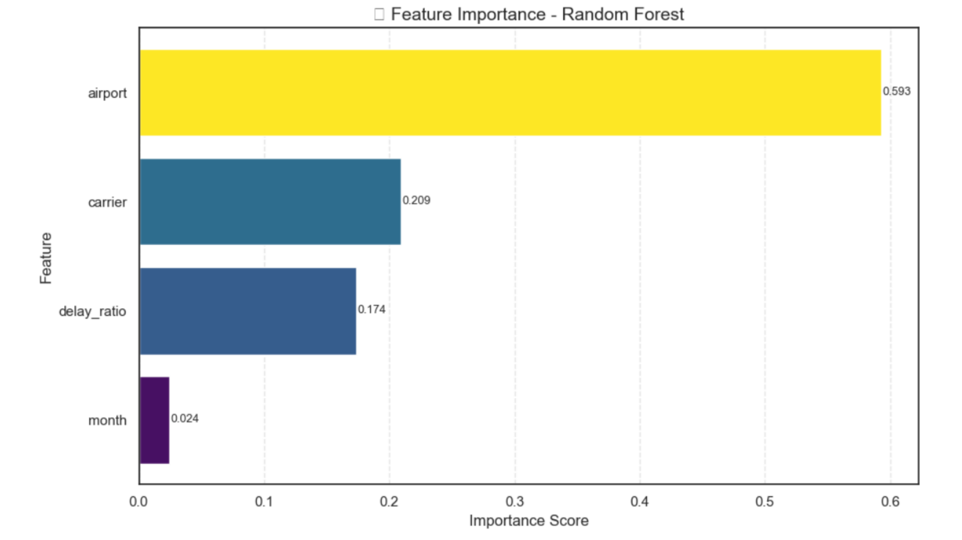
- **1** = Fully controllable (e.g., carrier operations)
- **0.5** = Partially controllable (e.g., airport congestion)
- **0** = Not controllable (e.g., time of year)

We then computed an **OAI (Operational Adjustability Index)**:

👉 OAI = Feature Importance × Adjustability Score

📈 Final Feature Rankings

Feature	Importance	Adjustability	OAI Score	Interpretation
Airport	0.593	0.5	0.297	Moderate control via gate assignment, decongestion
Carrier	0.209	1.0	0.209	High control via crew, aircraft scheduling
Delay Ratio	0.174	0.5	0.087	Useful planning metric; limited direct control
Month	0.024	0.0	0.000	Seasonal; not controllable



✈️ Final Recommendations — Data-Driven Delay Mitigation

- 🔧 **Operational Improvements**
• Introduce buffer time in scheduling to reduce cascading delays from late aircraft.
• Implement flexible crew and aircraft turnaround strategies to improve resilience during disruptions.

- 🌦️ **Weather-Adaptive Planning**
• Use machine learning models trained on historical weather-delay patterns to forecast high-risk windows.
• Pre-allocate resources and reroute flights in anticipation of severe weather events.

- 🛫 **Airspace & Airport Coordination**
• Strengthen coordination with ATC and FAA to manage congestion at NAS-impacted airports (e.g., EWR, JFK).
• Optimize gate and taxiway assignments at hubs like ORD, ATL to reduce ground delays.

- 📊 **Predictive Risk Scoring**
• Deploy route-specific delay risk models using historical cause-tagged data.
• Integrate real-time model predictions into airline operations dashboards for proactive decision-making.

- 📉 **Carrier-Level Adjustments**
• Recommend B6 (JetBlue) and WN (Southwest) reoptimize routes and aircraft allocation at peak-delay airports.
• Encourage investment in real-time scheduling systems for carriers with high delay ratios.

- 📈 **Strategic Delay Mapping (OAI Framework)**
• Prioritize mitigation for delay causes with high feature importance and high controllability (e.g., late aircraft over NAS).
• Use the Operational Adjustability Index (OAI) to guide investment in controllable delay factors.

Thank you!