Optimizing Air Travel: A Data-Driven Approach to Flight Delay Analysis and Prediction

ARPIT KUMAR
22117028

Project Objectives

01

Uncover Hidden Patterns

02

Develop Predictive Capability

03

Generate Actionable Insights

04

To help airlines minimize operational disruptions.

METHODOLOGY

I drop the columns with missing values because columns data depends on each other

Data Cleaning & Feature Engineering

Exploratory Data Analysis (EDA)

Performed data analysis on different columns

Balance the data , encode the categorial columns and applied different classification models

Classification (Delayed: Yes/No)

Regression (Delay Duration in Minutes)

Applied regression models on the data to know the expected average delay in minutes

By SHAPscore we can say that arr_flightsHeavily influences

Operational Adjustability Index(OAI) & SHAP

TOOLS USED : Python, pandas, matplotlib

MODEL USED : Linear and logistic regression , Random Forest Classifier and Regressor , XGBoost Classifier and Regressor

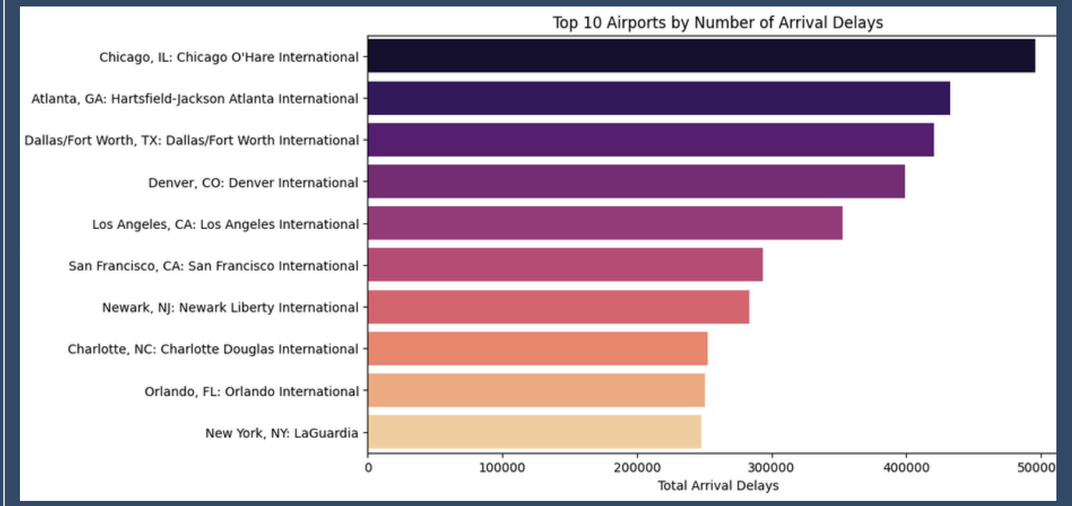
KEY FINDINGS

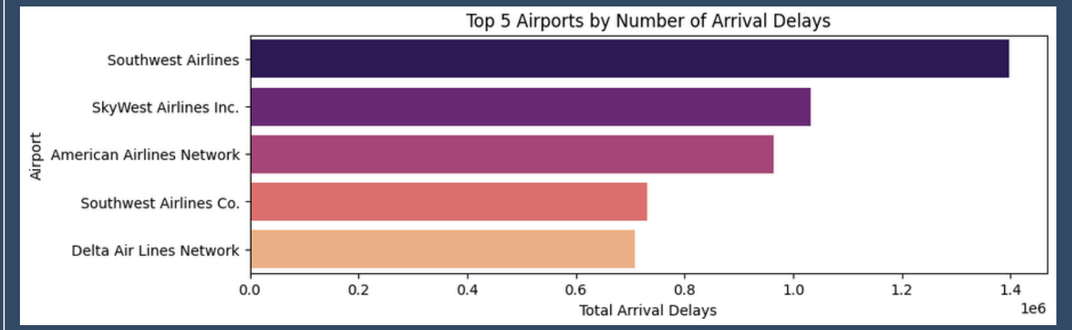
Patterns Identified:

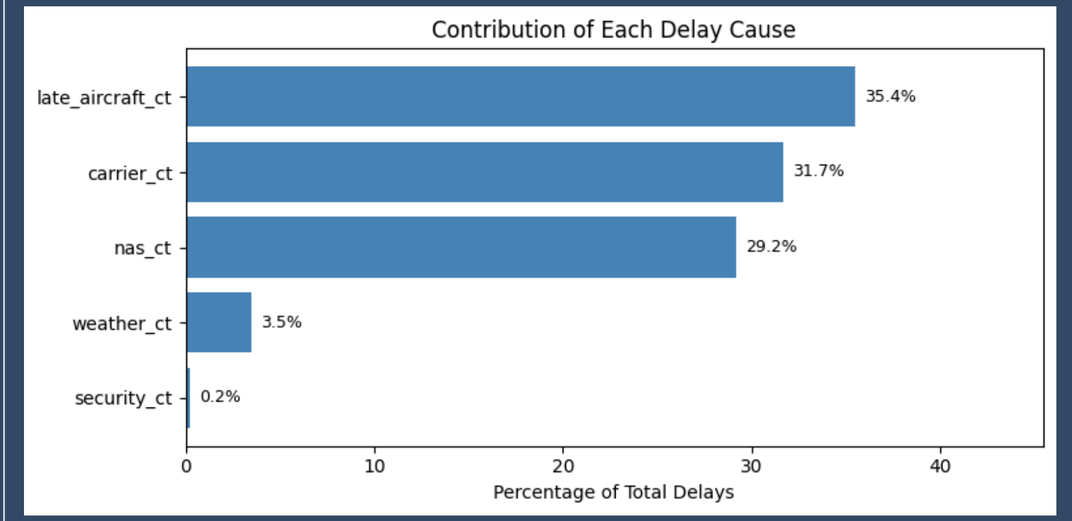
- Certain airports (e.g., Chicago, ATLANTA) show most delay
- SouthWest airline and Skywest airline show max delayed flights
- All airports are in major metros, suggesting delays stem from high traffic and complex networks.

Top Delay Drivers:

- Late Aircraft (35.4%) – Top delay cause; reflects tight turn arounds and poor schedule buffers.
- Carrier Delays (31.7%) – Airline-specific issues; opportunity for internal process improvement.
- NAS Delays (29.1%) – Air traffic/system inefficiencies; highlights need for ATC and airport upgrades.
- Weather + Security (3.75%) – Minimal impact overall, indicating smooth operations and favorable conditions.





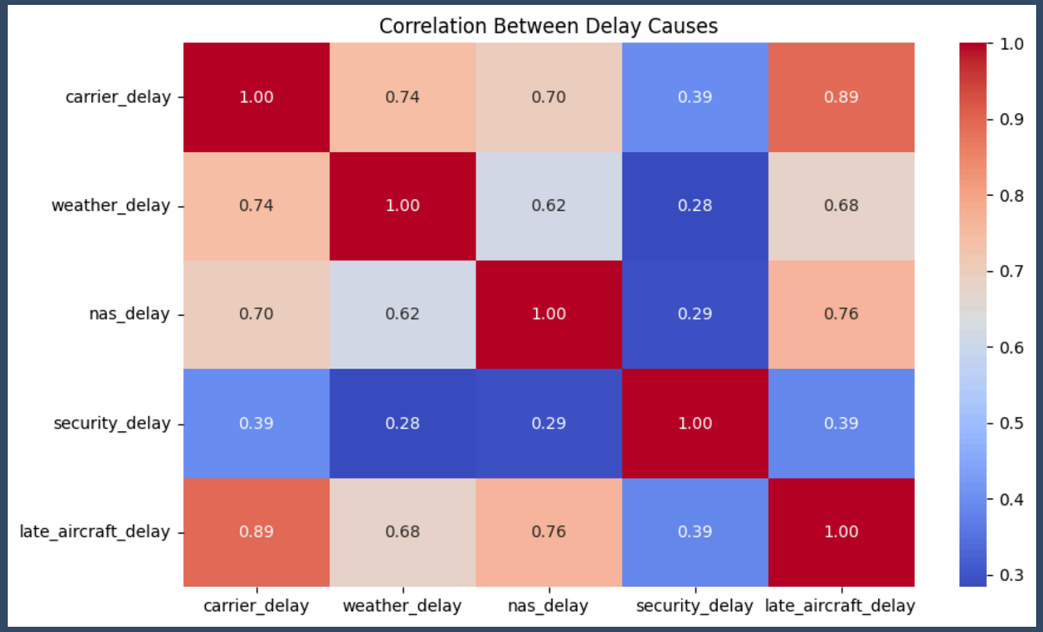


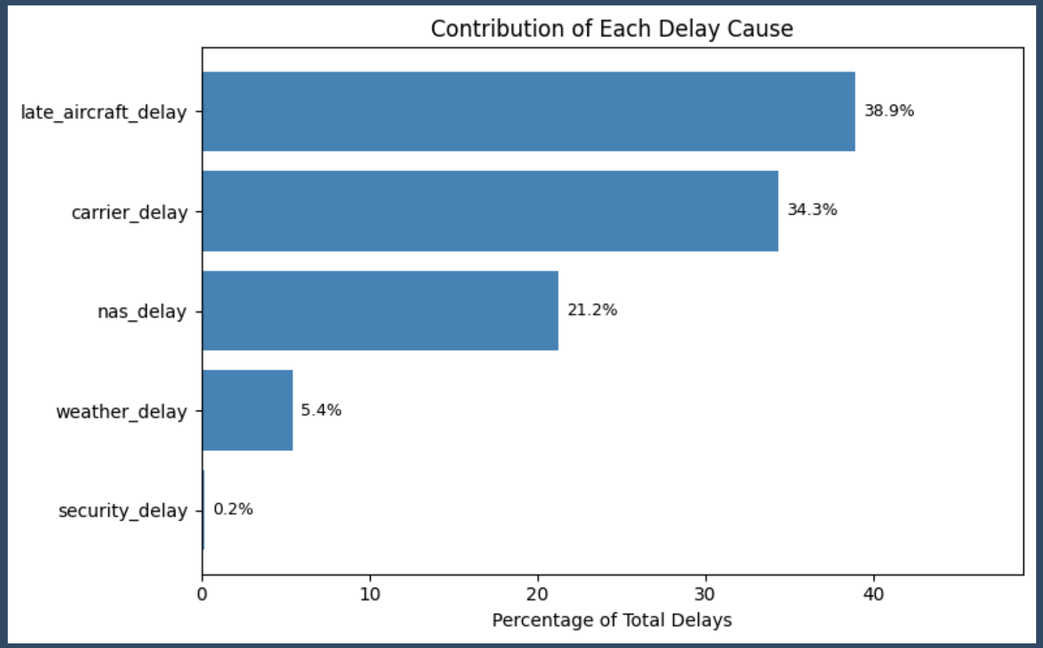
Correlation Between Delay:

- Late aircraft delays have the strongest correlations with other delay causes, especially carrier and NAS delays.
- Carrier delays are closely linked to both weather and system-related (NAS) issues.
- Security delays are rare and mostly independent of other causes.

Total Arrival Delay Distribution :

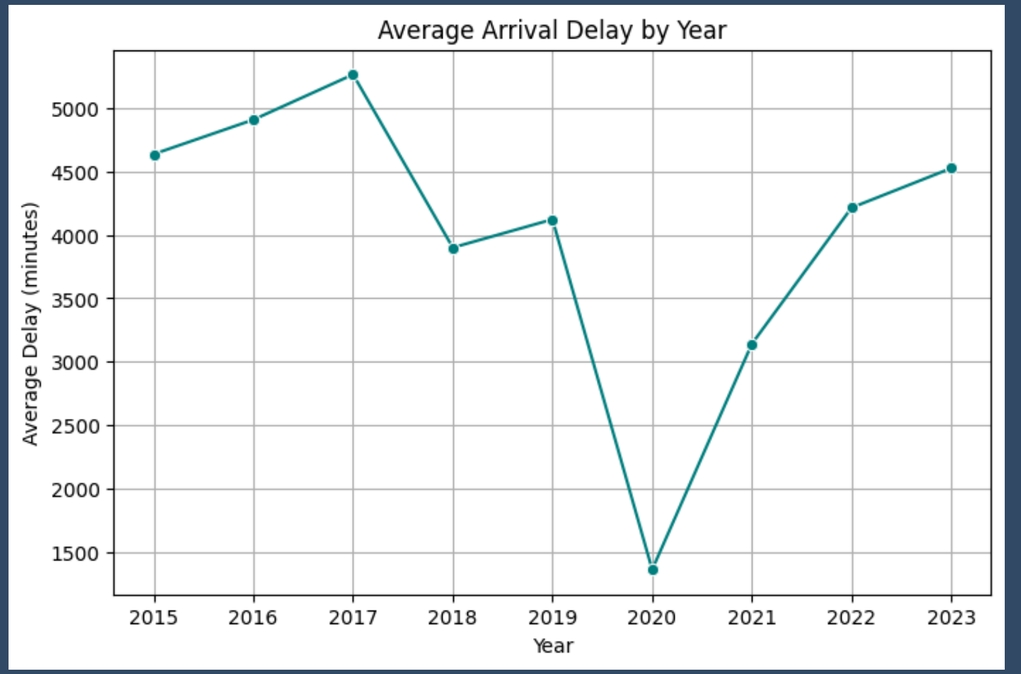
- Late Aircraft (38.89%) – Primary delay cause; due to previous flight arriving late, creating cascading delays.
- Carrier Delays (34.38%) – Airline-specific issues like crew, maintenance, or baggage handling inefficiencies.
- NAS Delays (21.24%) – Caused by traffic volume, congestion, and ATC-related inefficiencies.
- Weather + Security (5.49%) – Minor impact, likely due to favorable conditions and efficient safety protocols.





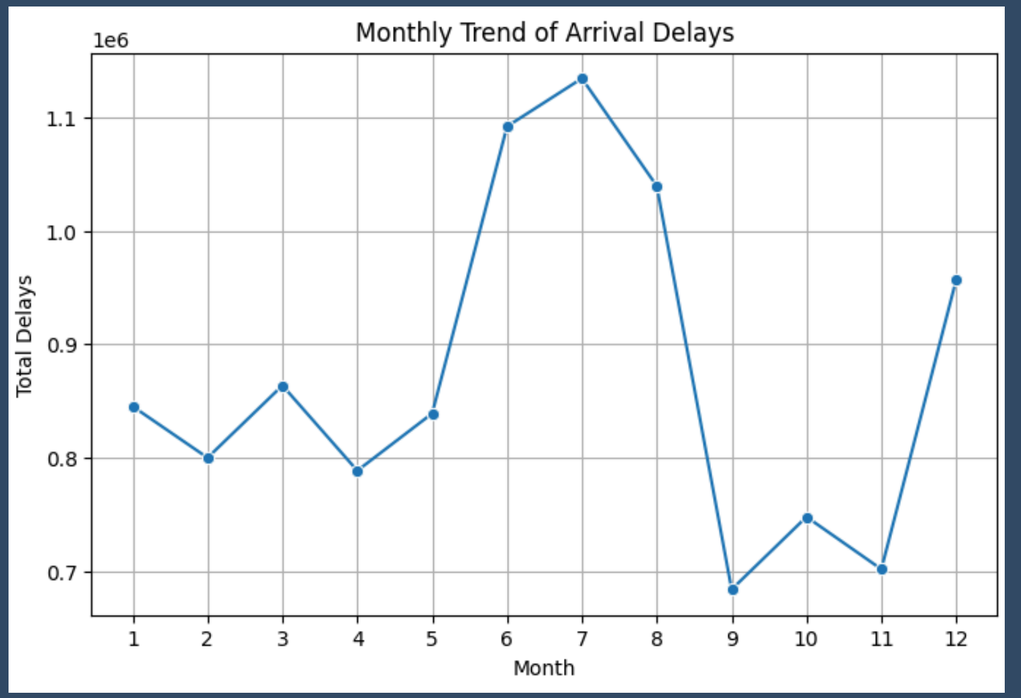
Average Arrival Delay by Year :

- Highest Delay: 2017 had the most delays — likely due to operational overloads.
- Lowest Delay: 2020 saw a sharp dropdue to the COVID-19 pandemic.
- Post-COVID Recovery: Delays climbed steadily from 2021 to 2023 (4400 mins) as flights resumed.



Total Delayed Flights by Month :

- Peak Delays: July, June, and August show the most delays — due to summer travel and monsoon impact.
- Lowest Delays: Sept, Nov, and Oct have the fewest delays likely due to lower demand and better weather.
- Moderate Months: January, March, and May show steady, moderate delays — indicating operational stability.
- Overall: Summer and year-end months are most delay-prone due to high travel and weather disruptions.



Model Performance

Classification Model (Delayed or Not): XGBoost Classifier Model is fitting best

- Accuracy: 96.88%

Regression Model (Minutes of avg_delay_per_delayed_flight):XGBoost regressor Model is fitting best

- MAE: 15.76 minutes (After removing large Outliers)
- RMSE: 24.12 minutes
- R2 Score: 37.3%

By SHAPscore we can say that arr_flightsHeavily influences predictions

suggests that arrival traffic volume at the airport is a major factor in predicting delays. Likely correlated with congestion.

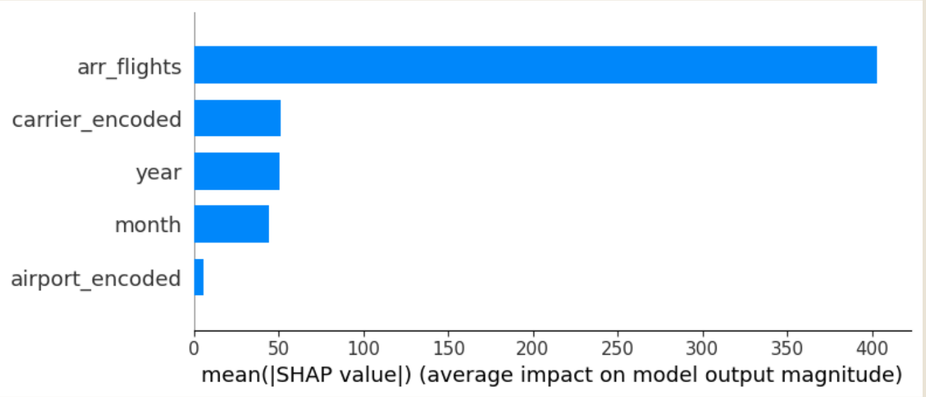
XGBoost Classifier Report:

	precision	recall	f1-score	support
0	0.75	0.50	0.60	1675
1	0.98	0.99	0.98	34075
accuracy			0.97	35750
macro avg	0.86	0.75	0.79	35750
weighted avg	0.97	0.97	0.97	35750

Confusion Matrix:
[[836 839]
[276 33799]]
✅ Accuracy: 96.88 %

Model Comparison Table

Model	MAE	RMSE	R^2
Linear Regression	21.12	30.22	0.016
Random Forest	16.58	25.06	0.323
XGBoost	15.76	24.12	0.373



Actionable Recommendation

1. Reduce Late Aircraft Turnaround Delays (Top Cause - ~39%)
    - Implement buffer times between flight arrivals and departures to absorb minor delays.
    - Improve gate allocation and ground operations to ensure faster turnaround.
2. Optimize Airline Internal Processes (Carrier Delays - ~34%))
    - Enhance crew scheduling systems to reduce unavailability.
    - Standardize maintenance procedures and reduce downtime via ML-based failure prediction.
3. Upgrade Air Traffic and NAS Coordination (NAS Delays - ~21%))
    - Collaborate with ATC authorities to optimize slot allocations and reduce congestion.
    - Invest in smarter scheduling during high-traffic months (May–August).
4. Seasonal Demand Management)
    - Increase staffing and fleet readiness during peak summer and holiday months.
    - Use historical delay trends to anticipate and mitigate seasonal congestion.
5. Resource Allocation:
    - Use SHAP scores + OAI to prioritize operational control efforts.

Thank **you!**