Flight Delay Analysis And Prediction

Introduction

Objective:

This project aims to analyze patterns in airline delays using real-world operational data, and to develop predictive models that can answer:

- Will a delay occur?
- If yes, how long will it last?

By combining exploratory data analysis with machine learning and explainability tools, the goal is to provide insights that help airlines reduce controllable delays and improve service quality.

Tools & Technologies:

- Python, Pandas, Matplotlib, Seaborn, Numpy, Plotly, Scikit-learn
- Random Forest Classifier, XGBoost, SHAP, OAI, Linear Regression, Logistic Regression
- SMOTE for imbalance handling

Dataset Summary:

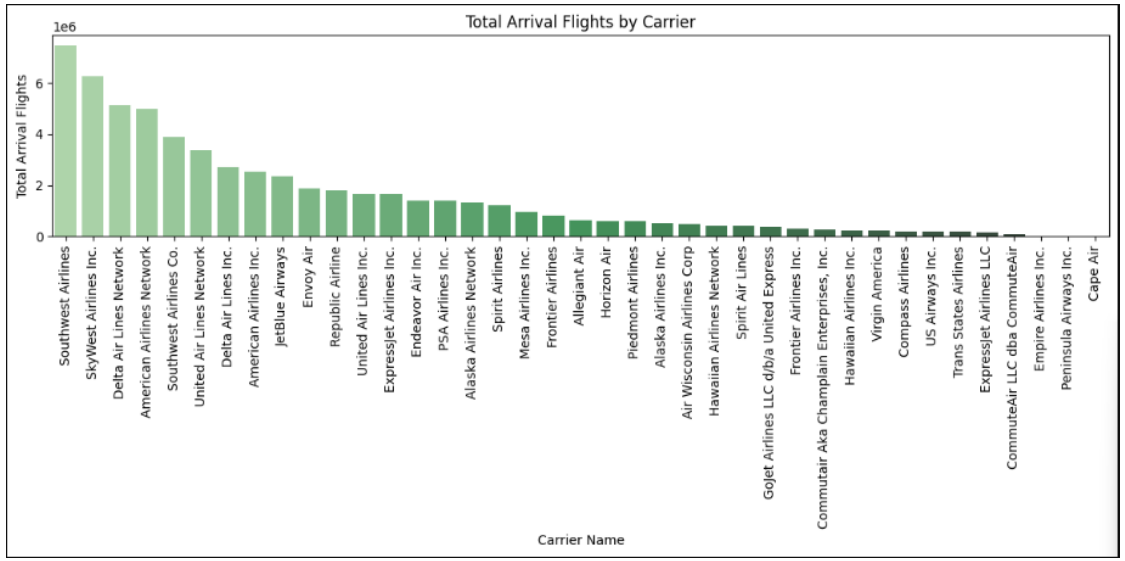
- ~179,000 records
- 21 features including carrier, airport, month, delay causes, and arrival times etc.

By:
Anuj Mittal
23117025

Exploratory Data Analysis

Flight Volume by Airport and Carrier:

- Atlanta, Chicago O'Hare, and Dallas Fort Worth are the top 3 airports by arrival volume.
- Southwest Airlines, SkyWest, and Delta handle the most flights across the dataset.
- These high-traffic nodes are likely to experience more delays simply due to volume pressure.
- Understanding which airports and carriers dominate the network helps in targeting delay reduction strategies.



Exploratory Data Analysis

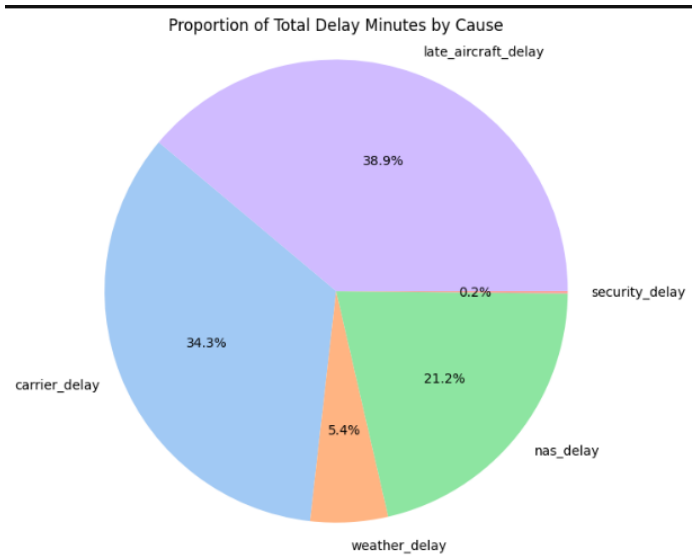
Late aircraft delays contribute the most to overall delay minutes (≈39%), followed by carrier delays (≈34%).

On average, delays due to late aircraft last the longest, followed by carrier and NAS-related delays.

Delays are highly correlated across types — e.g., late aircraft delay is strongly linked to both carrier and NAS delays.

Weather delays are frequent but generally shorter, and security delays are minimal overall.

This breakdown helps in identifying controllable vs uncontrollable delay categories — essential for operational planning.



Geographical Delay Trends

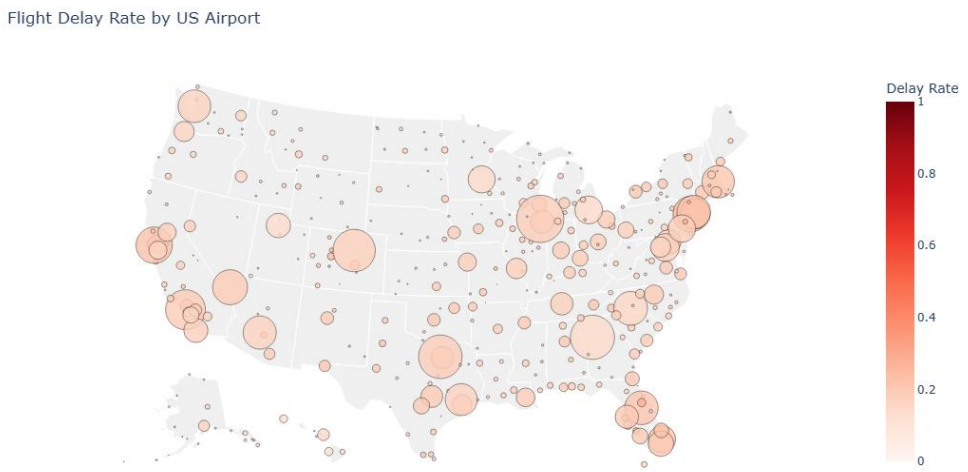
The chart highlights the top 15 airports with the highest flight delay rates.

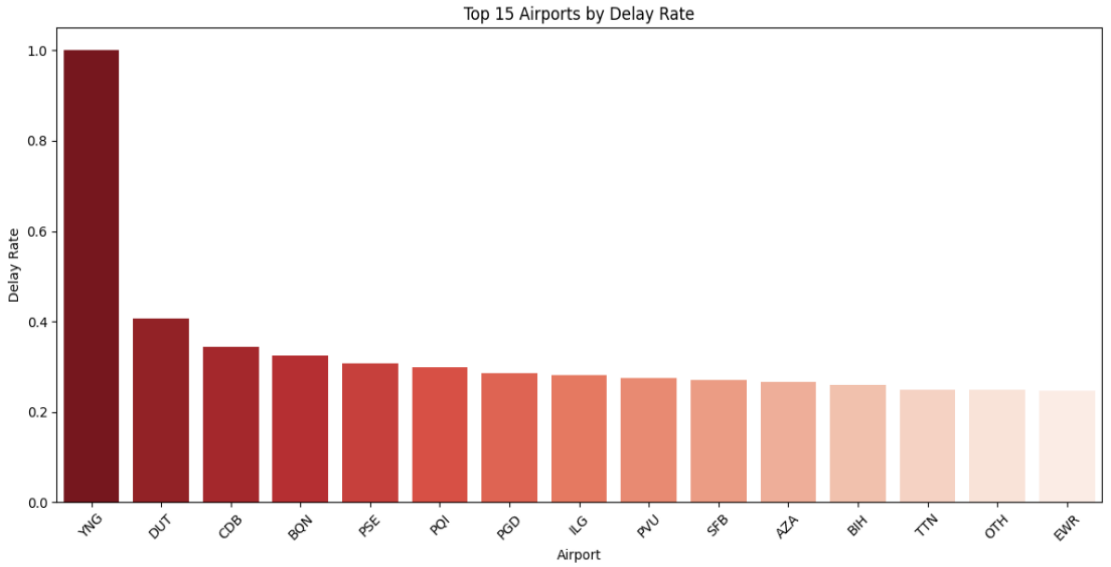
Youngstown (YNG), Dutch Harbor (DUT), and Cold Bay (CDB) top the list, with over 30–100% of flights delayed.

These airports may suffer from limited capacity, weather sensitivity, or operational constraints.

Delay rate is calculated as the percentage of arriving flights that were delayed by 15+ minutes.

Identifying such high-delay airports helps in prioritizing operational improvements and targeted resource allocation.





Classification Model – Will the Flight Be Delayed?

Objective:

To classify whether a flight group will be delayed by 15 minutes or more using calendar, airport, and delay patterns.

Target Variable:arr_del15

1 = Delayed (>=15 minutes)

0 = Not Delayed

Features Used:

month **:** To capture seasonal delay patterns.

arr_flights **:** Total flights for that carrier-airport-month.

late_aircraft_delay **:** Total delay due to late aircraft.

carrier_encoded, airport_encoded **:** Numerical codes for carrier and airport.

is_winter, is_summer, is_festive**:** Flags for seasonal and festival periods.

Models Used:

- Logistic Regression
- Random Forest
- XGBoost

Classification Model – Will the Flight Be Delayed?

Imbalance Handling:

- Dataset was heavily imbalanced (delayed flights were rare)
- Applied **SMOTE** to generate synthetic samples for the minority class

Best Results (Random Forest + SMOTE):

- Accuracy: **~89%**
- F1-score (Delayed class): **~0.86**
- Improved recall and balance after SMOTE application

Regression Model – How Long Will the Delay Be?

Objective:To predict the delay duration (in minutes) in case a flight is delayed.

Target Variable:arr_delay – Delay time in minutes

Features Used: (Same as classification)

month, arr_flights, late_aircraft_delaycarrier_encoded, airport_encodedis_winter, is_summer, is_festive.

Models Used: Linear Regression

XGBoost Regressor

Best Performance (XGBoost):

MAE: ~759 mins

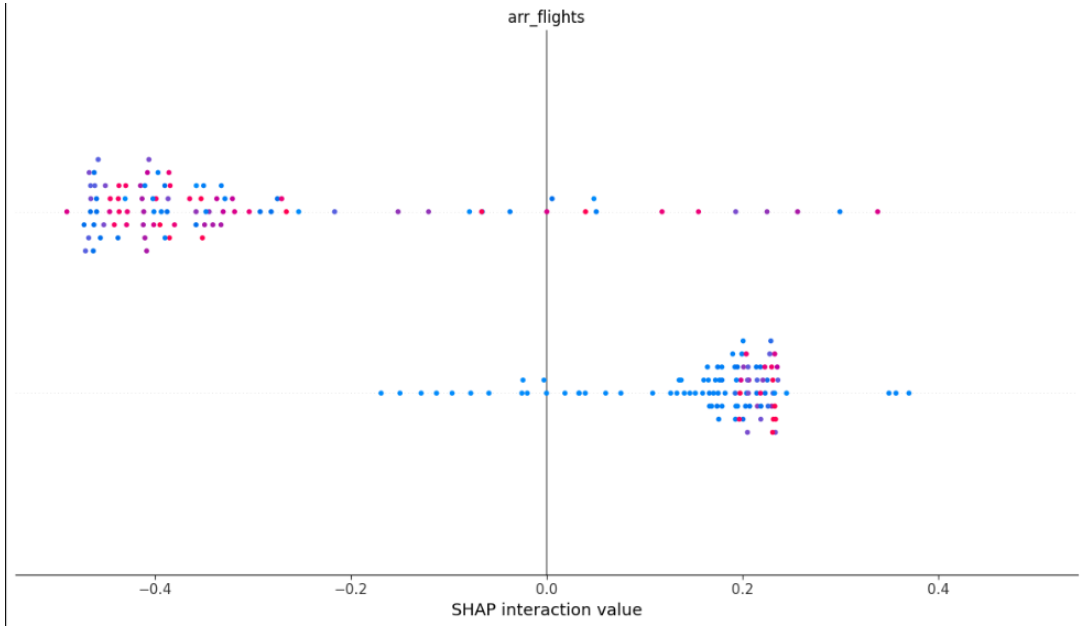
RMSE: ~3552 mins

R^2 Score: ~0.92

Insight: XGBoost captured delay variability more effectively, especially in high-delay scenarios. The model can help estimate delay length, aiding in resource planning and passenger communication.

SHAP

- To make the classification model explainable, we applied SHAP (SHapley Additive Explanations).
- SHAP revealed that features like arr_flights, carrier_ct, and late_aircraft_ct had the highest influence on the model's predictions.
- This insight helps stakeholders understand why a delay was predicted — not just the result.



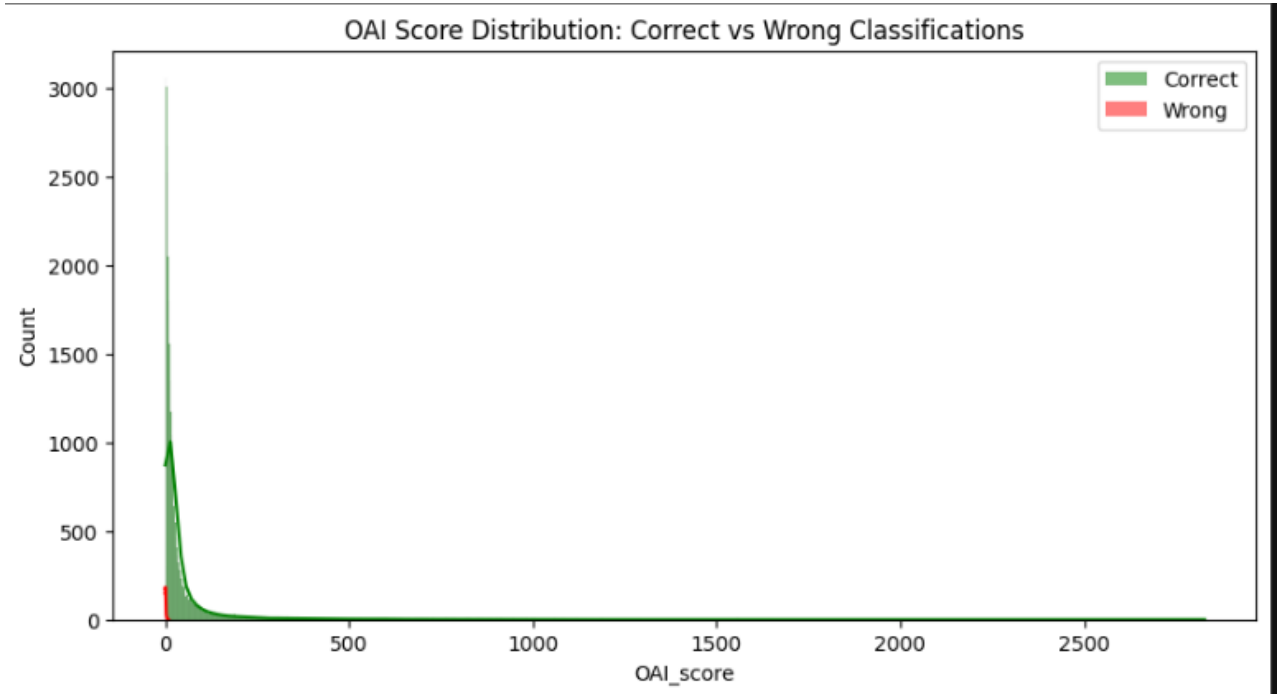
OAI – Operational Adjustability Index:

- We calculated an OAI score to prioritize delays that airlines can actually control, such as:

carrier_ct, late_aircraft_ct, security_ct ✅

Less weight to: weather_ct, nas_ct ❌

- This ensured our model was action-oriented, not just predictive.



Thank You