Optimizing Air Travel: A Data-Driven Approach to Flight Delay Analysis and Prediction
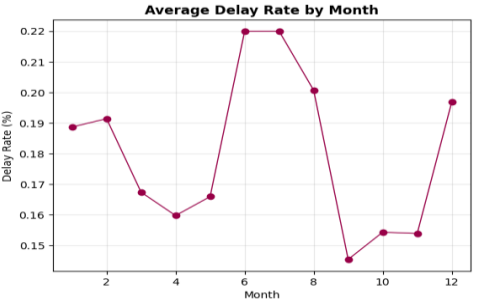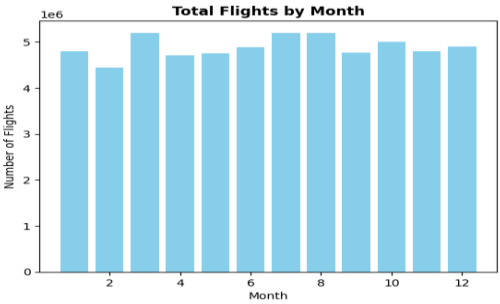
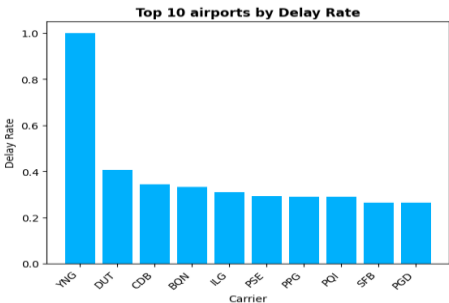Meenakshi
22116053
ECE(4y)

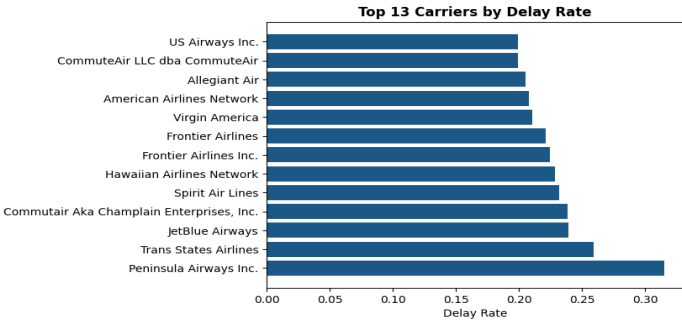# EXPLORATORY DATA ANALYSIS

Track Monthly Delay Trends
Visualize how average delays change over time.
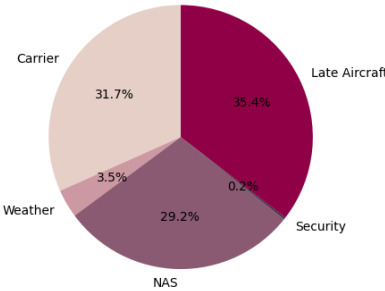
Identify Top Airlines & Airports
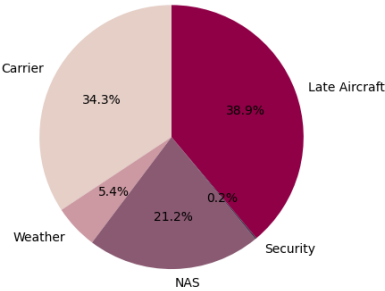Pinpoint carriers and airports most affected by delays.

Understand Seasonal Impact
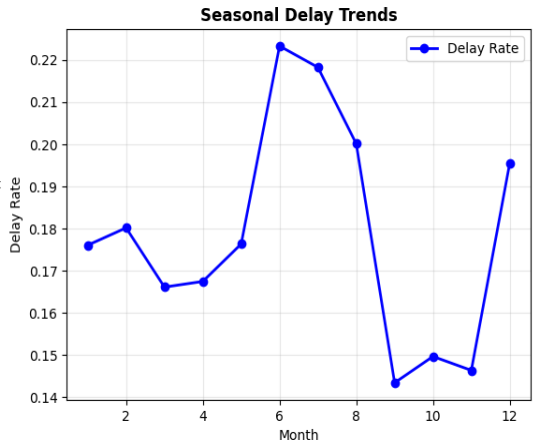Assess how seasonal changes influence delay frequency.

Conduct Root Cause Analysis
Examine specific delay types: carrier, weather, NAS, security, and late aircraft.

Advanced Delay Insights

Average Delay by Cause:
Bar chart showing mean delay contributions

from:

Correlation Heatmap:
Reveals interrelationships among different delay factors.

Delay Rate Distribution:
Histogram showing how delay rates vary across all flights.

Summary Statistics:
Displayed key metrics like mean, median, min, max, and standard deviation of delay rates.
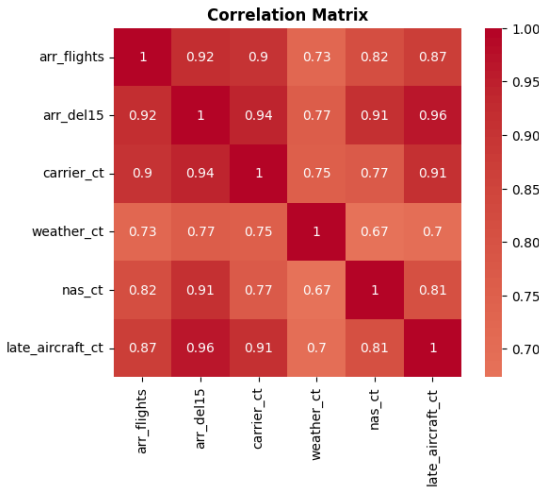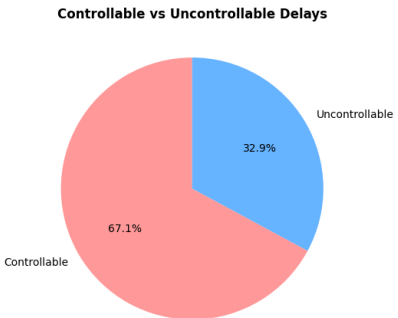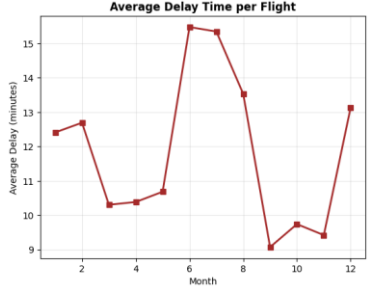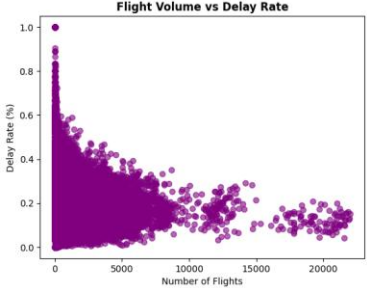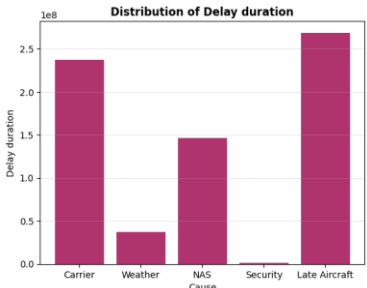
SEASONAL PATTERNS:
- Month with highest average delay: July (0.22%)
- Month with lowest average delay: September (0.15%)

OVERALL STATS:
    Total flights analyzed: 58,676,070.0
    Delayed flights: 10,494,073.0
    Overall delay rate: 17.88%

CARRIER PERFORMANCE:
- Most delay-prone carrier: Peninsula Airways Inc. (0.32%)
- Most reliable carrier: Cape Air (0.08%)
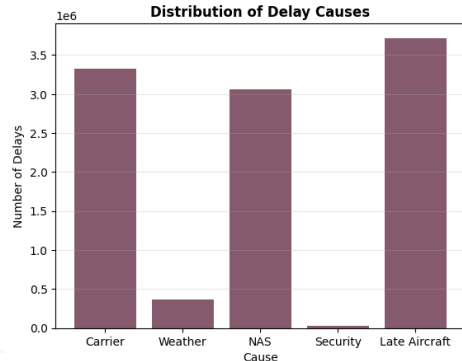
Feature Engineering

1. Core Performance Metrics
- `delay_rate`: Proportion of delayed flights
- `avg_delay`: Average delay per flight (in minutes)

2. Delay Cause Proportions
- `carrier_prop`: Delay share due to airline
- `weather_prop`: Delay share due to weather

3. Target Variables
- `is_delayed`: Binary label for delay presence
- `target_delay`: Average delay for regression

4. Historical Reliability Features
- `carrier_his`: Historical delay rate per airline
- `airport_his`: Historical delay rate per airport

5. Encoded Categorical Variables
- `carrier_encoded`: Encoded airline names
- `airport_encoded`: Encoded airport names

Cleaning & Filtering:

Removed rows with missing values in key delay-related columns.

Excluded records where arr_flights = 0 to avoid invalid calculations.

Outliers in continuous variables such as avg_delay and delay_rate were treated using the **IQR (Interquartile Range) method**, ensuring robust model performance without removing valid extreme delay instances.

Operational Adjustability Index (OAI)

What is OAI?
A custom metric to measure how much of the delay is operationally controllable.

Formula
OAI = (carrier_delay + late_aircraft_delay) / total_delay

It helps airline ops teams prioritize improvements on controllable delays.

High OAI → Focus on scheduling, fleet readiness, crew management

Low OAI → Indicates external issues like weather/NAS delays

Classification model for Flight Delays

- For the classification problem , **Random Forest classifier** is used after balancing data using **SMOTE**
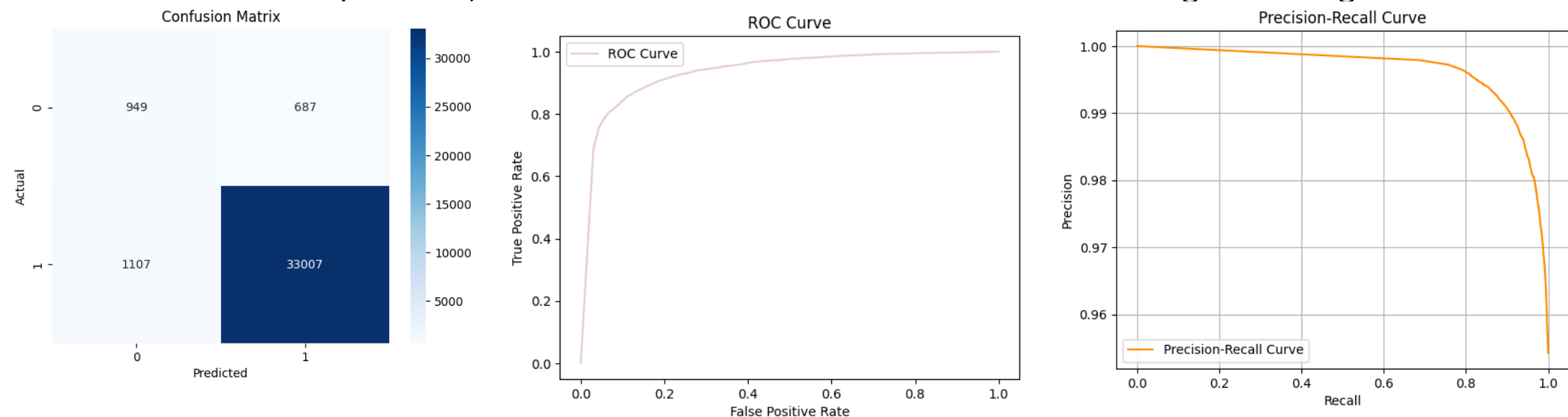
Evaluation Metrics:

```
Accuracy: 0.9498181818181818
Precision: 0.9796106131655488
Recall: 0.9675499794805652
F1 Score: 0.973542944785276
ROC AUC: 0.9346907349197204
Confusion Matrix:
[[  949   687]
 [ 1107 33007]]
Classification Report:
              precision    recall  f1-score   support

           0       0.46      0.58      0.51      1636
           1       0.98      0.97      0.97     34114

    accuracy                           0.95     35750
   macro avg       0.72      0.77      0.74     35750
weighted avg       0.96      0.95      0.95     35750
```
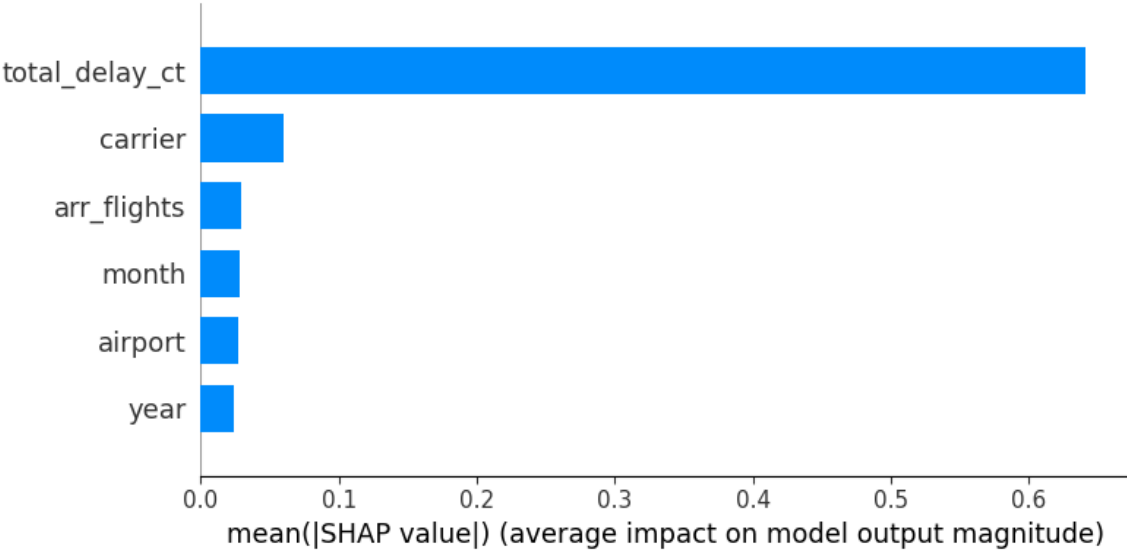
Predicting Arrival Delays with different Models

Linear regression model
• Mean Absolute Error (MAE):7.20
• Mean Squared Error (MSE): 236.01
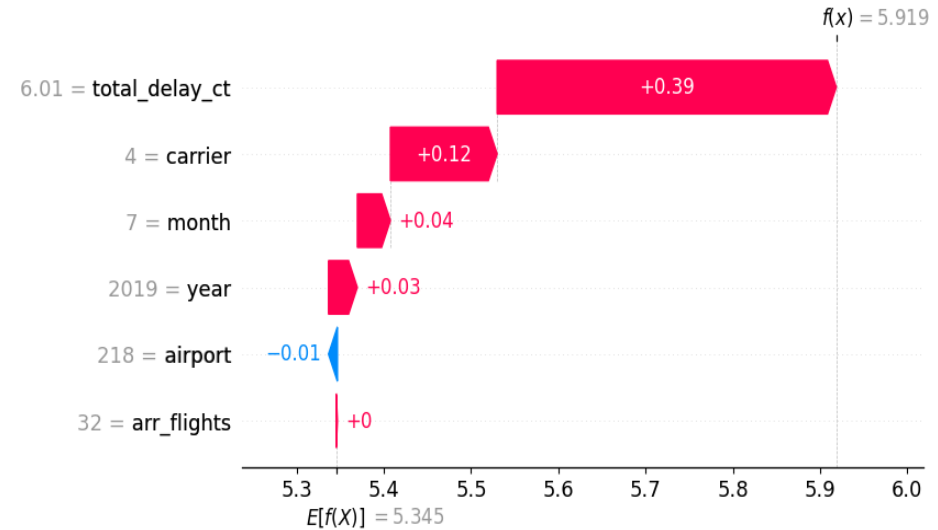
XG BOOST Regression model
Mean Absolute Error (MAE): 67.65 minutes
Root Mean Squared Error (RMSE): 91.80 minutes
R2 Score: 0.71

Actionable Recommendations & Consulting Insights

1
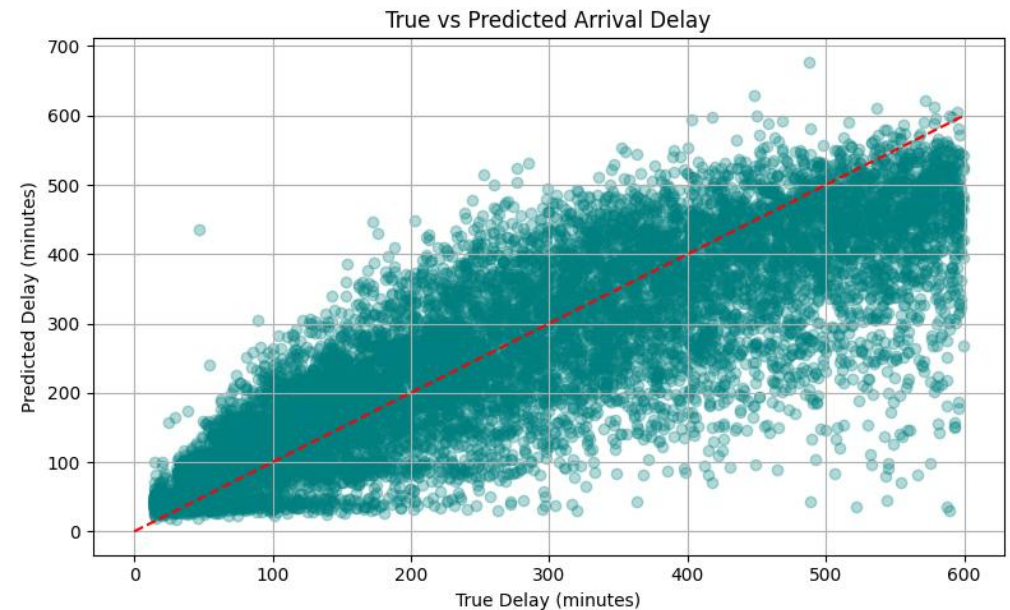
Improve Ground Operations for Controllable Delays

- Focus on reducing **carrier-related** and **late aircraft delays**, as these are the most controllable and contribute significantly to total delay time.
- Implement tighter turnaround management and buffer times in scheduling.

2

Schedule Adjustments and Network Planning

- Reschedule flights during peak congestion hours or airports with **historically high delays**.
- Use insights from seasonal features (e.g., more delays during summer) to plan better staffing and resource deployment.

3

Prioritize Airports and Carriers with High Delay Rates

- Use the **carrier_historical_performance** and **airport_historical_performance** metrics to identify consistently underperforming segments.
- Initiate targeted interventions like performance audits or collaborative improvement initiatives.

4

Enhance Real-Time Communication & Customer Engagement

- Improve **proactive communication** with passengers during expected NAS/weather-related delays.
- Deploy digital tools (SMS/Email/Apps) for early alerts and rebooking options.

5

Data-Driven Resource Allocation

- Allocate more operational resources (staffing, gates, crew) during **holiday/winter peaks** or **high OAI score periods**.
- Monitor **OAI-weighted duration** to dynamically adjust schedules.