SocBiz Open Project 2025

Data Analytics and Machine Learning Model

Topic-Flight Delay Prediction & Optimization using ML

Made by - Lakshya Gupta

Enrollment number - 24410015

Geologicaal Technology

Overview:

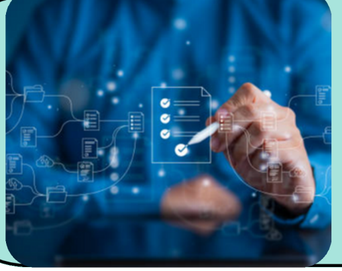
In the more complicated and time-sensitive era of air travel, airlines are experiencing growing operational challenges as a result of flight delays. Flight delays not only impact customer satisfaction but also cause substantial financial and logistical losses for airlines.

The Solution: *Data-driven insights*.
Using past flight data, we used machine learning to forecast the likelihood of a delayed flight and the extent of delay. We created the Operational Adjustability Index (OAI) to rank controllable delays and utilized SHAP explainability to make model predictions clear and actionable.

Important Strategies:

1. Predict & Prevent: Apply predictive models (84% accurate) to predict delays and act early.
2. Target Controllable Causes: Prioritize operational efforts on carrier and late aircraft delays (OAI Score: 29.92).
3. Act with Insight: Use SHAP-based dashboards to identify and respond to the most influential delay drivers per flight.

Why it Matters:
By addressing the most actionable causes of delays, airlines can enhance punctuality, lower operational expenses, and optimize the passenger travel experience.
This initiative allows airlines to progress from reactive delay management to proactive delay avoidance, resulting in a more intelligent, effective air transport system.



Methodology :

1 Data Preparation

- Created delay_time_per_flight & delay_ratio_per_flight as target metrics.
- One-hot encoded carrier and airport; selected only numeric fields.

2 EDA

- Peak delays in summer and winter months.
- Carrier delay and Late aircraft delay are the most frequent controllable causes.
- Delay types visualized using heatmaps, bar charts, and time plots.

3 Model Building

- Built XGBoost Classifier (81.6% acc) & Regressor (MAE: 12.2 min).
- Split data (80/20), tuned models with GridSearchCV.
- Evaluated using accuracy, F1, MAE, RMSE, and R^2.

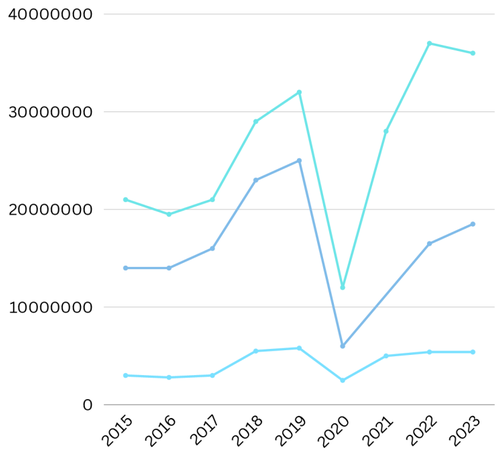
4 OAI Integration

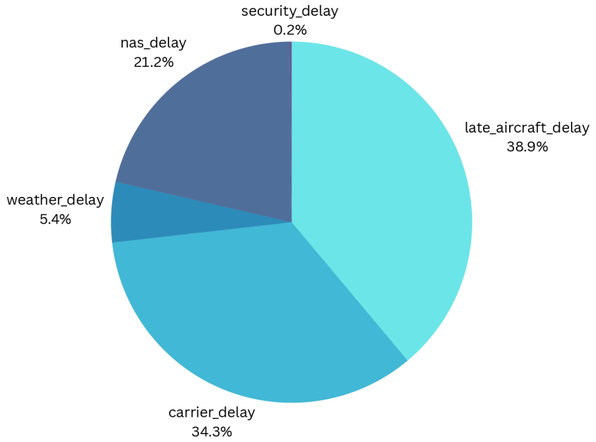
- Weighted controllable delays (3×) to create Operational Adjustability Index.
- OAI score = 29.92 → ~30% delays are controllable.
- Focused on delay types airlines can control directly.

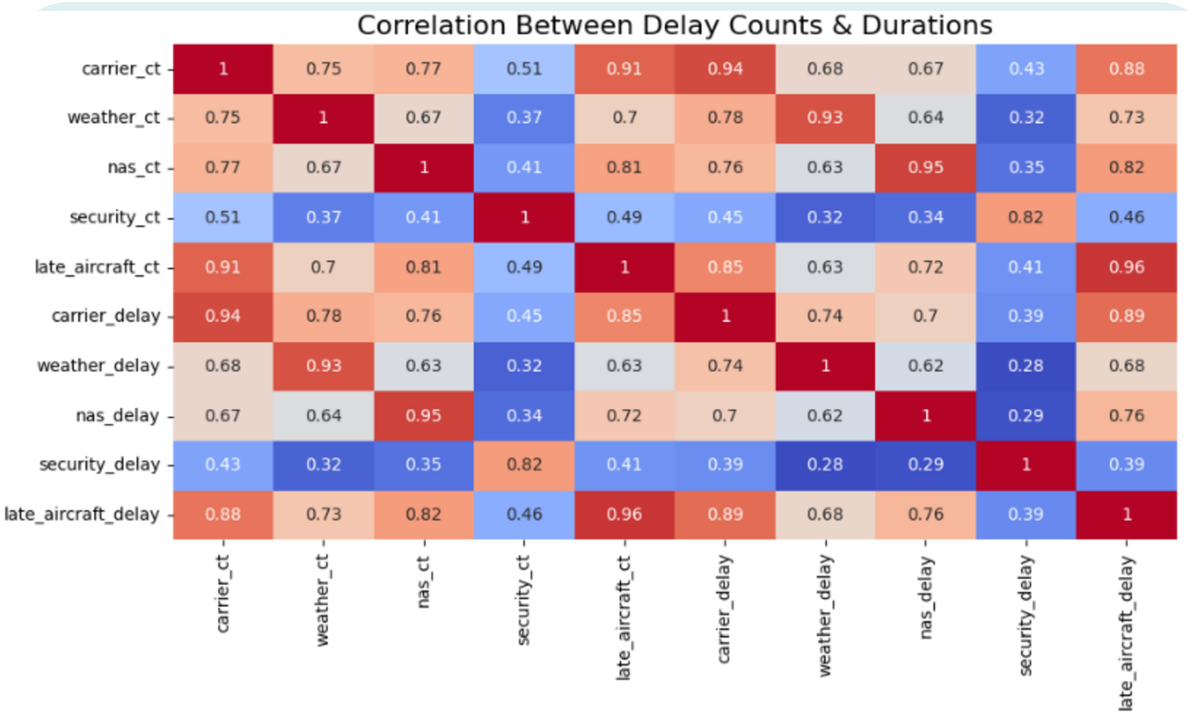
5 SHAP Analysis

- Used SHAP to explain delay drivers at flight level.
- Top drivers: carrier_AA, airport_DFW, month, arr_flights.
- Visualized insights using SHAP summary and force plots.

Highlights from EDA:

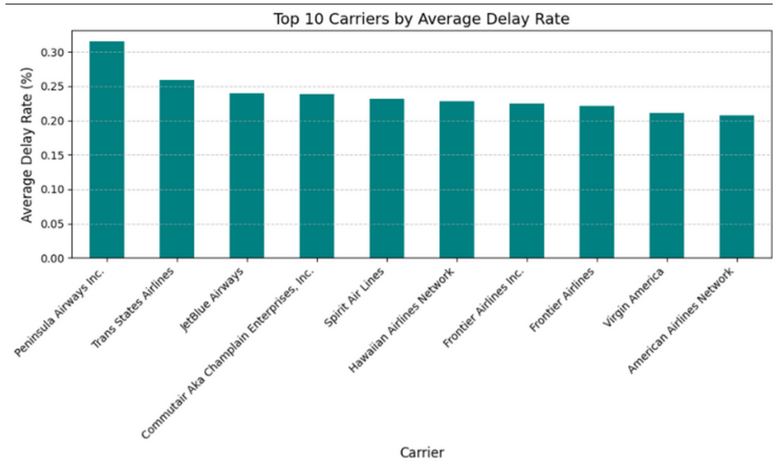






Carrier delays have surged post-2020, becoming the dominant and steadily increasing cause of total delay time across recent years.

Nearly 73% of total delays are caused by carrier and late aircraft issues, highlighting significant opportunities for airline-driven operational improvements

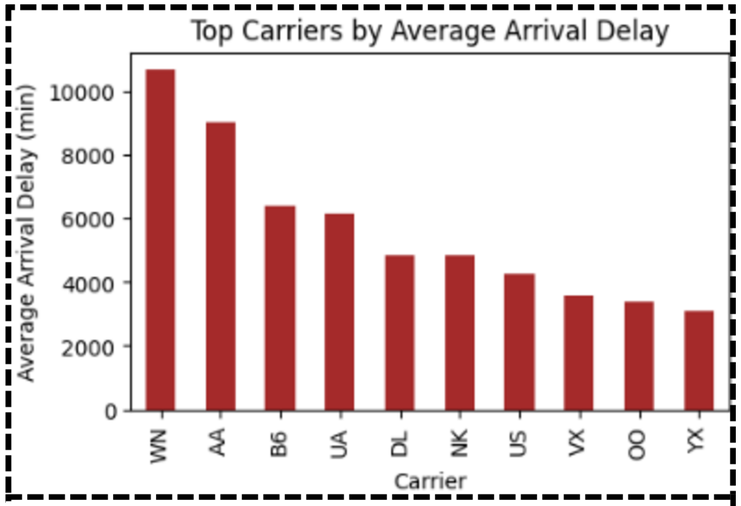


- Peninsula Airways and Trans States Airlines have the highest average delay rates, signaling consistently poor on-time performance.
- These carriers may require targeted operational reviews or resource adjustments to improve reliability and reduce chronic delays.

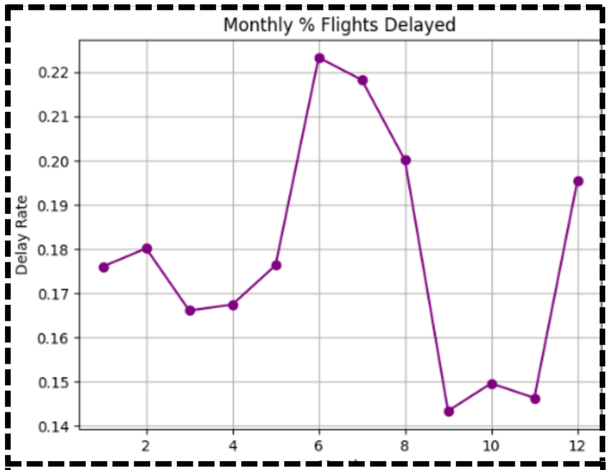
Insights:

- Carrier and late aircraft delays show the strongest correlations between count and duration (≥0.9), indicating they are consistent and impactful when they occur.
- These delay types are controllable by airlines, making them high-priority targets for operational improvement.
- NAS and weather delays also show moderate correlations, but are less tightly linked, reflecting variability in external conditions.
- Overall, the heatmap highlights that addressing carrier-related inefficiencies can significantly reduce total delay time across the network.

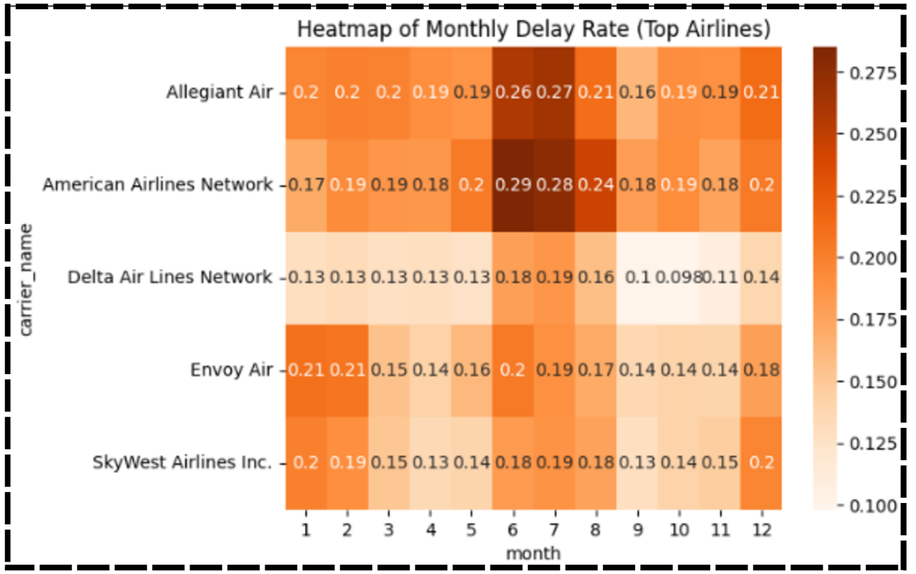
Airline Delay Analysis – Key Takeaways



Southwest Airlines Co. has the highest average arrival delay, suggesting major inefficiencies in its flight operations.

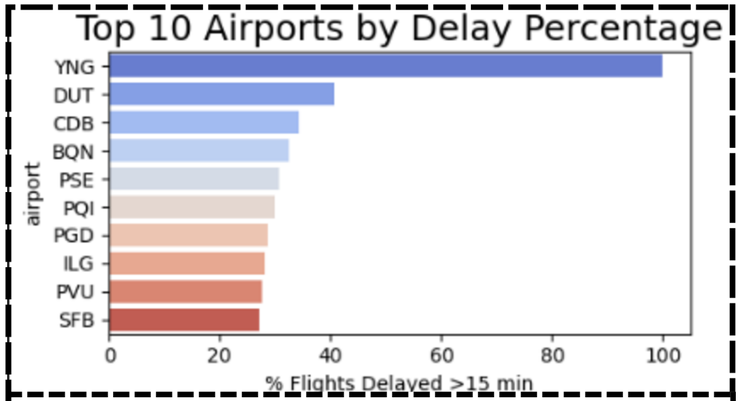


Flight delays peak in June and July, indicating summer months experience the highest disruption rates.

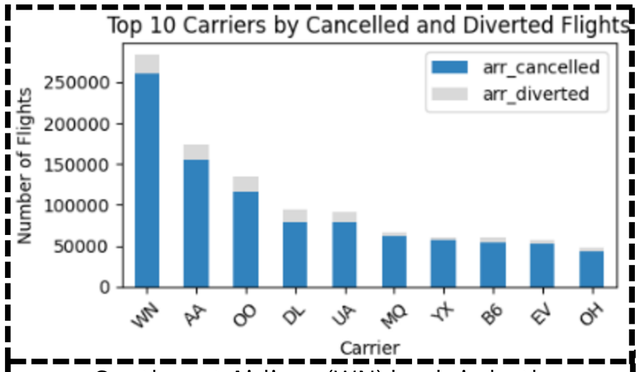


Insights:

- June and July show peak delay rates across most airlines, highlighting summer as the most delay-prone season.
- American Airlines Network experiences the highest single-month delay spike (0.29) in June, indicating capacity or scheduling strain.
- Delta and SkyWest consistently maintain lower delay rates year-round, reflecting stronger operational reliability.
- Overall, monthly delay patterns reveal that both seasonality and carrier-specific factors drive variability in punctuality.



Youngstown-Warren Regional Airport has the highest flight delay percentage, with nearly all flights experiencing delays over 15 minutes.



Southwest Airlines (WN) leads in both cancellations and diversions, followed by American Airlines (AA) and SkyWest (OO), indicating significant operational disruptions.

Model Performance

```python
def get_season(month):
    if month in [12, 1, 2]:
        return 'Winter'
    elif month in [3, 4, 5]:
        return 'Spring'
    elif month in [6, 7, 8]:
        return 'Summer'
    else:
        return 'Fall'
df_new['season'] = df_new['month'].apply(get_season)
```

You are engineering a new feature called "season" from the existing "month" column:
For example: Summer may have more weather delays (storms), Winter may have snow-related issues, etc.

```python
df_new['delay_time_per_flight'] = df_new['arr_delay'] / df_new['arr_del15']
df_new.dropna(subset=['delay_time_per_flight'], inplace=True)
df_new['delay_ratio_per_flight'] = df_new['arr_del15'] / df_new['arr_flights']
df_new.dropna(subset=['delay_ratio_per_flight'], inplace=True)
```

This code creates two new columns to calculate average delay per delayed flight and delay ratio per flight, and removes any rows with missing values from these calculations.

```python
#Encoding(convert alphabhet into numerical)
data2_encoded = pd.get_dummies(df_new, columns=['carrier', 'airport','season'],
                               drop_first=True, dtype=int)
data2_encoded = data2_encoded.select_dtypes(include=[np.number])
```

This converts categorical columns (carrier, airport, season) into numeric one-hot encoded variables and retains only numeric columns for modeling.

```python
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

It standardizes the training and test feature data by scaling them to have zero mean and unit variance using StandardScaler.

```python
from sklearn.preprocessing import LabelEncoder,StandardScaler
from sklearn.linear_model import LinearRegression,LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier
from sklearn.metrics import precision_score, recall_score, f1_score
```

Trying different machine learning models on the data to find the most appropriate for the given test data

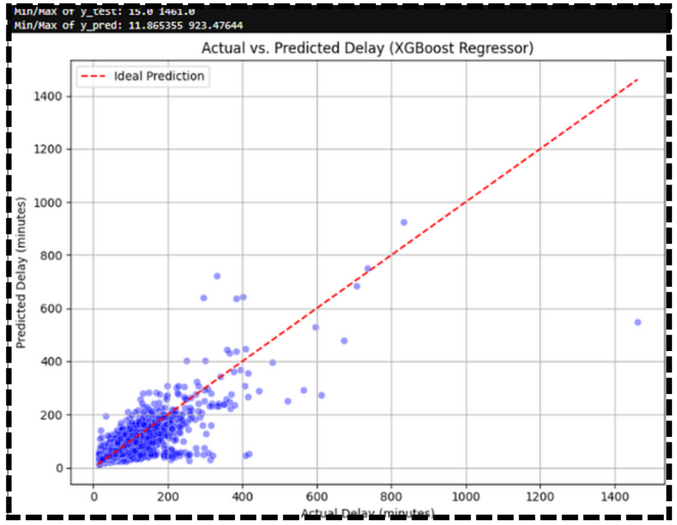
Regression

```
# Feature matrix and target vector
X = data2_encoded[feature_columns]
y = data2_encoded['delay_time_per_flight']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Creates feature matrix X and target y (delay_time_per_flight) for regression, then splits data into training and test sets.

```
RMSE: 21.66
MAE: 12.22
R² Score (Accuracy): 0.63
```

Visualizes regression performance metrics — lower MAE and RMSE values indicate good prediction accuracy.



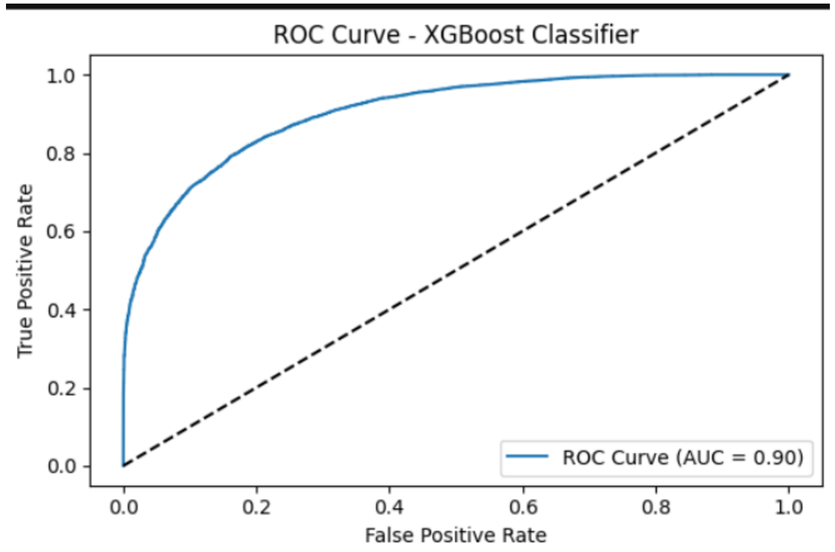
The model predicts short-to-moderate delays reasonably well,

Classification

```
data2_encoded['high_delay_ratio'] = (data2_encoded['delay_ratio_per_flight'] > threshold).astype(int)
X_cls = data2_encoded[feature_columns]
y_cls = data2_encoded['high_delay_ratio']
X_train, X_test, y_train, y_test = train_test_split(X_cls, y_cls, test_size=0.2, random_state=42)
```

Creates a binary classification target high_delay_ratio based on a threshold and splits the data for classification modeling.

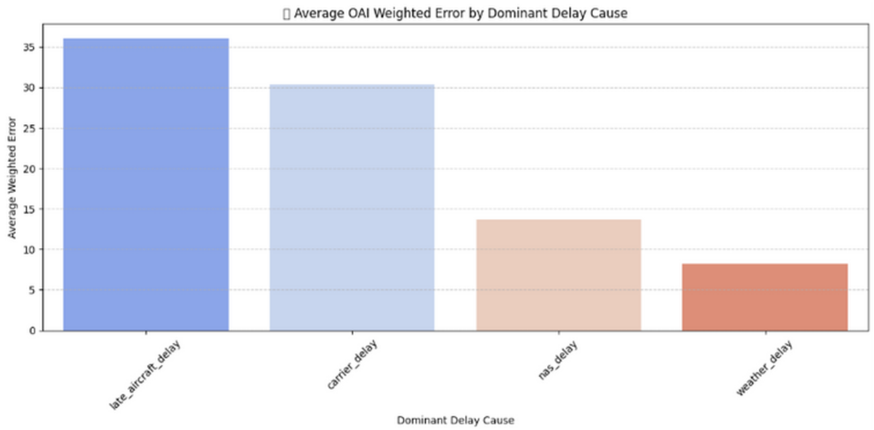
```
🔹 XGBoost Results
Best Params: {'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 100}
Accuracy : 0.8157539929475213
Precision: 0.8226674298797939
Recall   : 0.781936887921654
F1 Score : 0.8017852161785216
```



Displays the ROC curve for the XGBoost Classifier with AUC = 0.90, indicating strong binary classification performance.

OAI AND SHAP ANALYSIS

OAI:



Purpose:

OAI focuses model evaluation on controllable delays (like late_aircraft_delay and carrier_delay) by assigning them higher weights

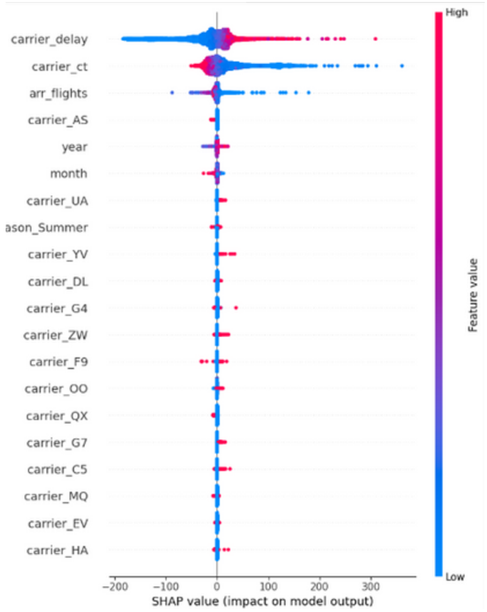
Actionability:

Helps prioritize flight operations where human or system intervention can reduce delay.

Insight from Chart:

- Late aircraft delay and carrier delay contribute the most to average OAI-weighted error.
- These two causes should be top priorities for delay reduction strategies.

SHAP:



Purpose:

- SHAP explains how each input feature increases or decreases the predicted delay for a specific flight.
- It assigns a value to each feature showing its individual contribution to the final prediction.

Top features:

- carrier_delay has the strongest impact on model output (both positive and negative).
- carrier_ct (carrier control time) and arr_flights also significantly affect predictions.

Interpretation:

- Red dots (high values) on the right show that high delay values lead to higher predicted delays.
- Blue dots (low values) on the left show the opposite — low delays reduce prediction scores.

Recommendation:

Actionable Recommend-ations

Prioritize High OAI Flights:
Focus on flights with frequent late_aircraft and carrier_delay, as they are the top contributors to controllable delays (OAI insight).

Tailor Interventions by Carrier:
Features like carrier_AS, carrier_UA, and others show specific delay behavior — develop carrier-specific strategies instead of one-size-fits-all approaches.

improve Operational Planning for High-Volume Airports:
Airports like Atlanta, Chicago O'Hare, and DFW handle the most flights — optimizing operations here will yield large-scale delay reductions.

Mitigate High-Delay Airports:
Airports like Youngstown-Warren and Unalaska have delay rates >90% — rerouting or infrastructure review may be needed.

Predict & Prevent Delay Clusters:
Use top SHAP features to anticipate systemic delay chains, e.g., when high carrier delay is paired with high arr_flights, and deploy dynamic rescheduling alerts.

Reschedule During Peak Delay Months:
June–August and December show the highest % delays — apply schedule buffering and preemptive staffing during these months.

Target High-Cancellation Carriers:
Peninsula Airways and Empire Airlines have the highest cancellation rates — collaborate or audit to reduce flight disruptions.

Thank You