Society of Business

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

DEVANSHI
devanshi@ch.iitr.ac.in

Society of Business

OPTIMIZING AIR TRAVEL WITH MACHINE LEARNING

Predicting Flight Delays & Identifying Controllable Causes

Problem

Air travel delays are a persistent issue impacting cost, operations, and customer experience.

- 1 in every 5 flights in the US is delayed
- Flight delays cost airlines $8.3 billion annually (source: FAA)
- Passengers lose over 28 million hours per year waiting
- Traditional analytics fail to separate controllable vs uncontrollable delays

Tools & Techniques

- Python (Pandas, Seaborn, Matplotlib)
- Scikit-learn (Random Forest)
- SHAP (Explainability)
- Custom metric: Operational Adjustability Index

Our Approach

Cleaned & explored data, trained models to predict delays and durations, and applied SHAP + OAI to generate airline-focused insights.

Project Objectives

- Predictive

Will a flight be delayed? (Binary classification)

- Quantitative

Estimate delay duration (Regression)

- Explainable

Use SHAP to understand why the model predicts a delay

- Actionable

Introduce OAI — a metric that gives priority to delays airlines can control (like carrier & late aircraft)

“21% of flights are delayed. What if we could predict & reduce the controllable ones?”

METHODOLOGY: FROM DATA TO DECISIONS

A Step-by-Step Data Science Pipeline

Data Collection

U.S. Department of Transportation: 170k+ flight records with delay reasons

Icon: Database / Excel / API icon

Exploratory Data Analysis

Visualized delay patterns by time, cause, airport; identified key bottlenecks

Icon: Bar chart / Magnifying glass

Modeling

Trained classification & regression models using Random Forest

Icon: Brain / gear / decision tree

Explainability (SHAP)

Used SHAP values to interpret model predictions and highlight key features

Icon: Puzzle piece / SHAP logo / lightbulb

Operational Index(OAI)

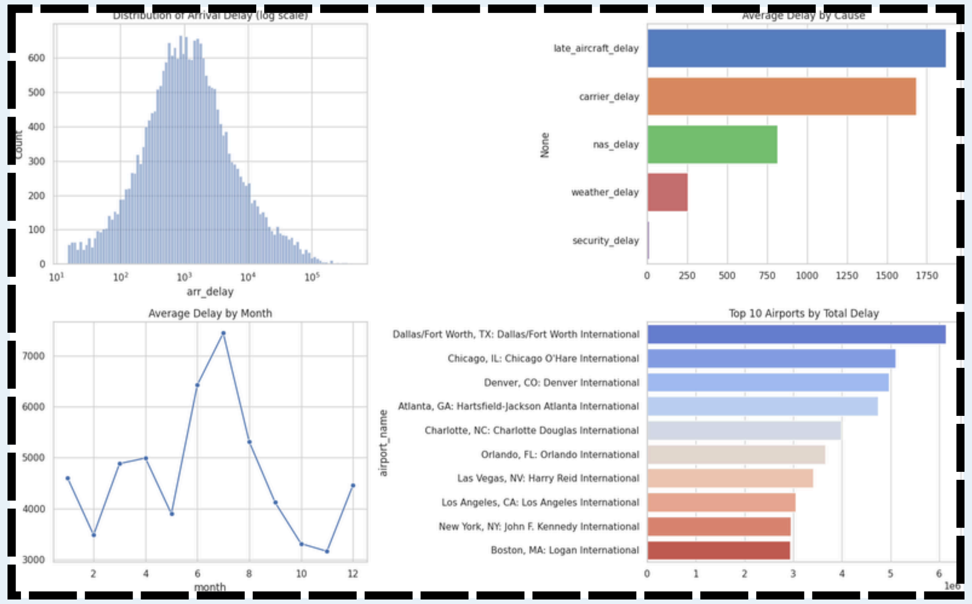
Designed a metric to prioritize controllable delays like carrier or aircraft issues

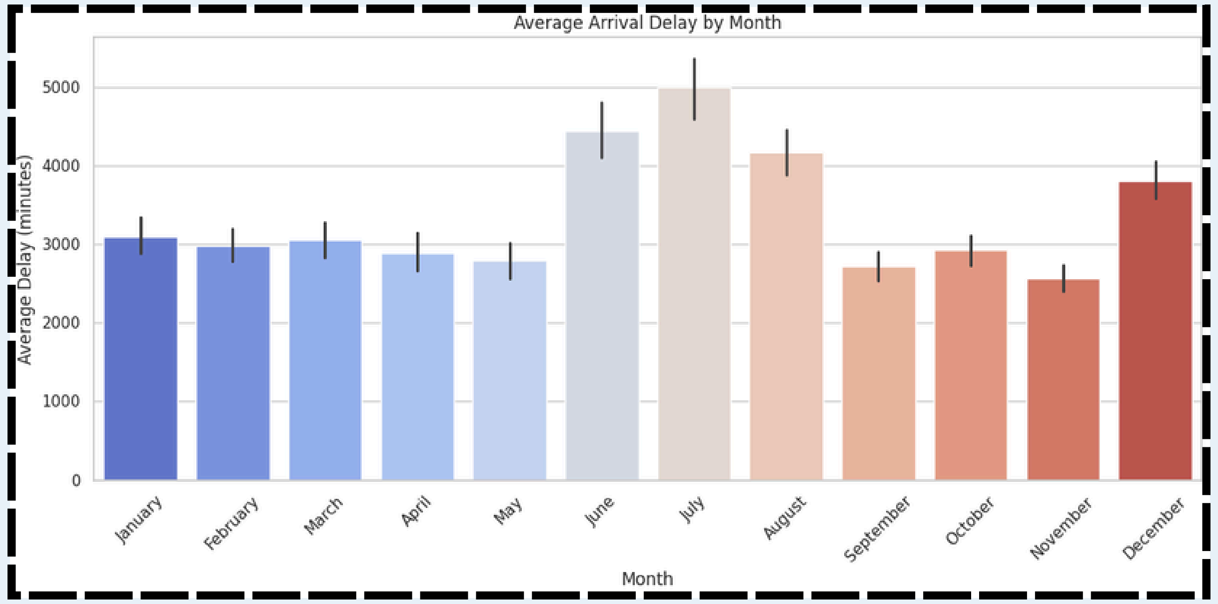
Icon: Target / control sliders / checklist

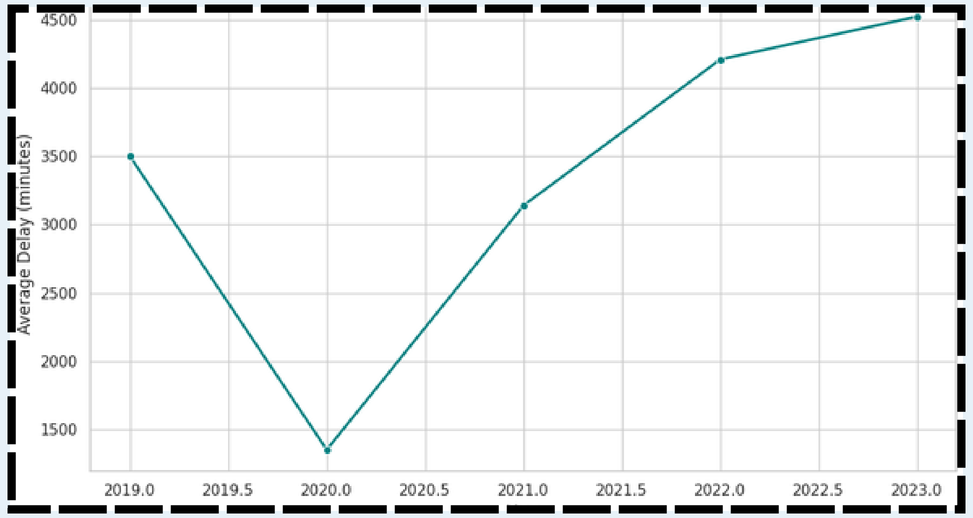
Society of Business

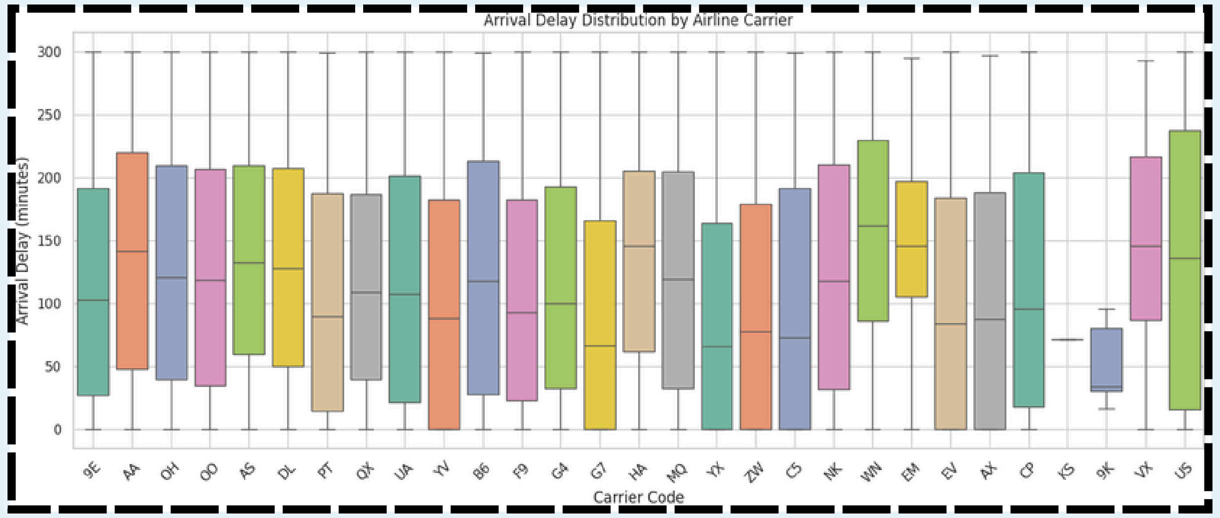
VISUAL INSIGHTS FROM EDA

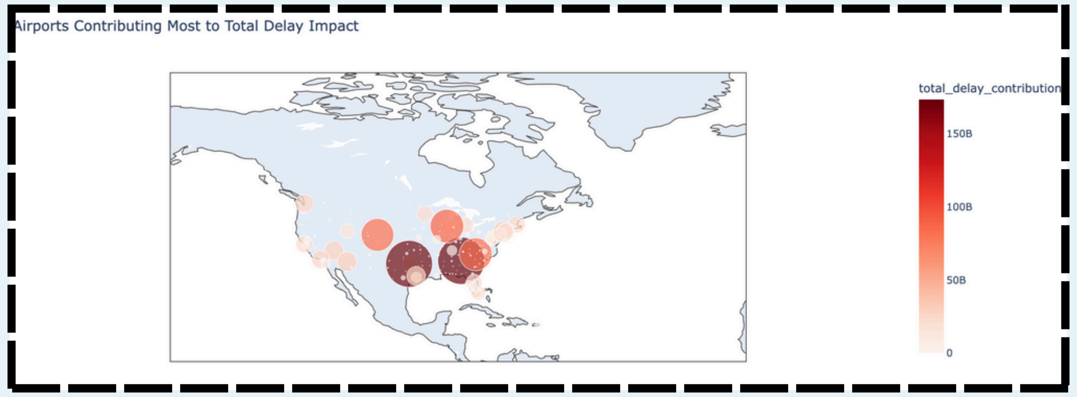
What the data revealed before modeling











Key Insights

- Late aircraft and carrier delays are the dominant causes
- Evening and peak-season flights face more delays
- Major hub airports (e.g., ATL, ORD, DFW) show the highest delay totals
- Delay patterns show strong monthly/seasonal variation
- Strong correlation between arr_flights and arr_delay
- Delays vary by month, peaking during holiday and summer seasons.
- Carrier-level analysis shows which airlines are more delay-prone.
- Geographical analysis highlights key airports that contribute disproportionately to system-wide delays.

Society of Business

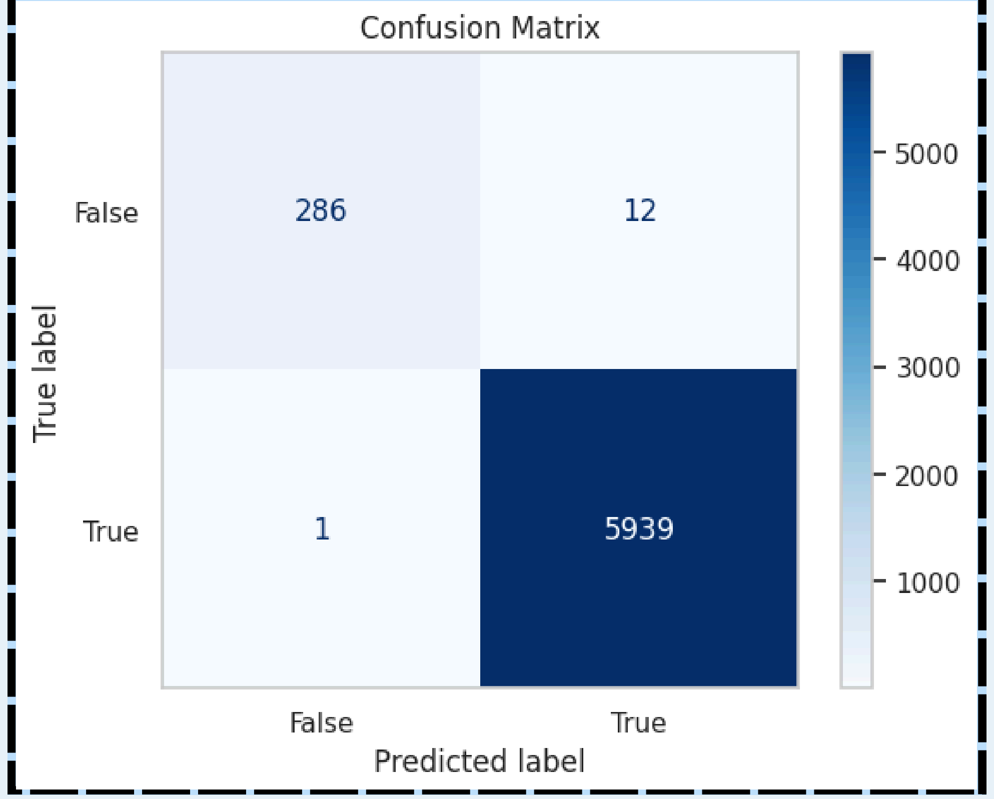
PREDICTING FLIGHT DELAYS WITH MACHINE LEARNING

Evaluating model accuracy & reliability

Classification Model

Task: Will the flight be delayed?
Model: Random Forest Classifier
Metrics:

- Accuracy: 99.79%
- Precision: 99.82%
- Recall: 99.64%
- F1-Score: 99.73%

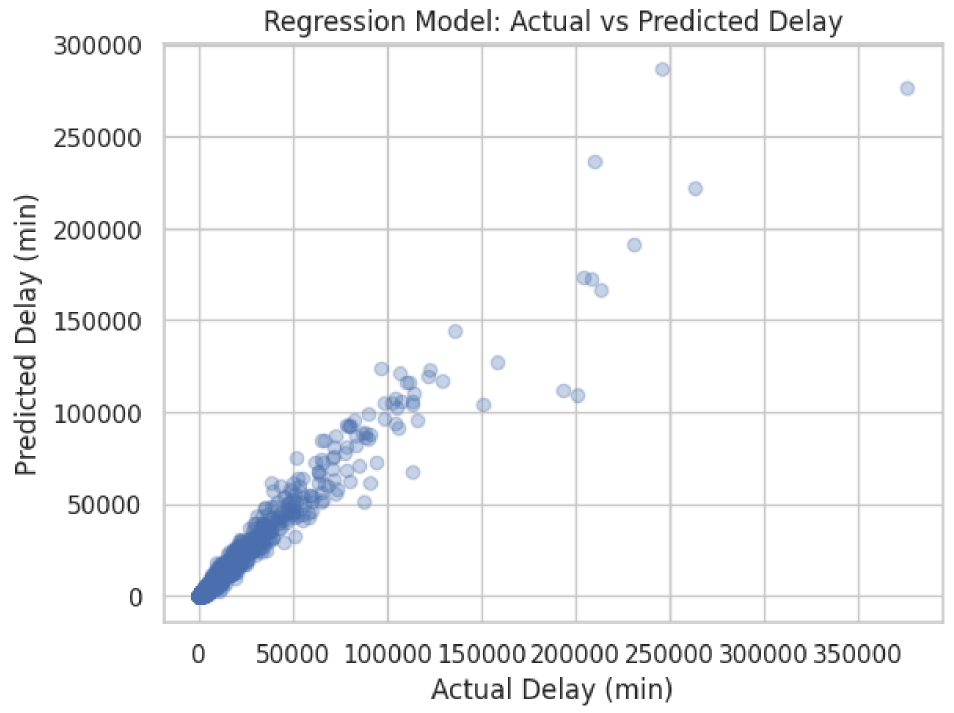


Regression Model

Task: How many minutes will the flight be delayed?
Model: Random Forest Regressor
Metrics:

- MAE: 800.41 minutes
- RMSE: 3159.76 minutes

The model performs best for short to moderate delays, with some overprediction for extreme cases



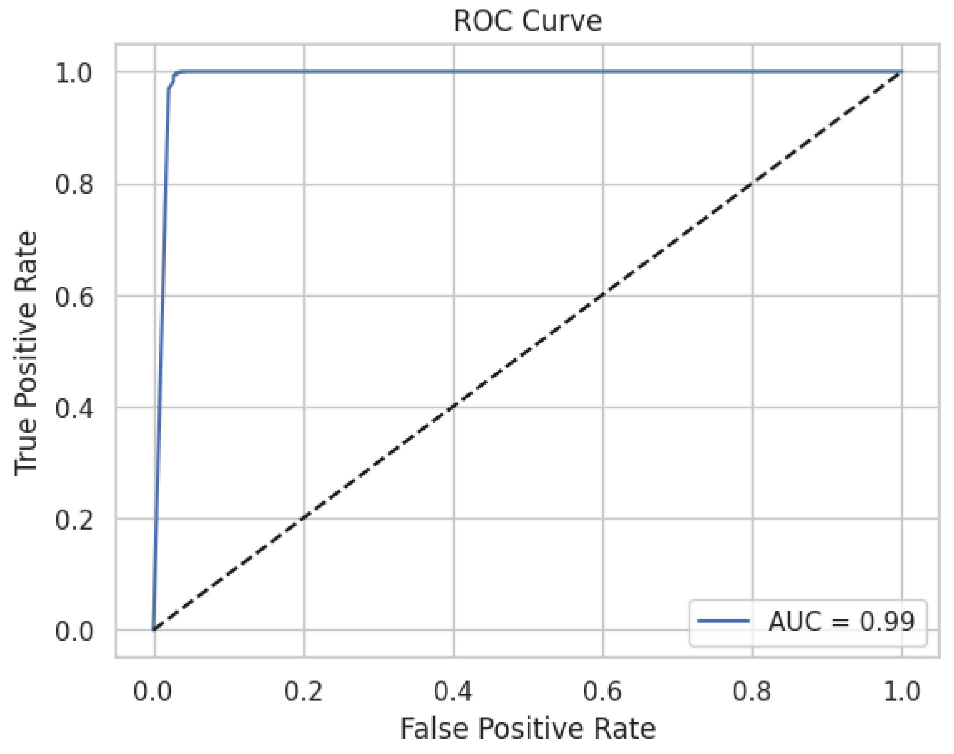
ROC Curve (Classifier Evaluation)

Excellent ROC Performance

- The classifier's ROC curve demonstrates outstanding separation.

AUC (Area Under Curve): 0.99

- This indicates the model is highly capable of distinguishing between delayed and on-time flights with minimal error.



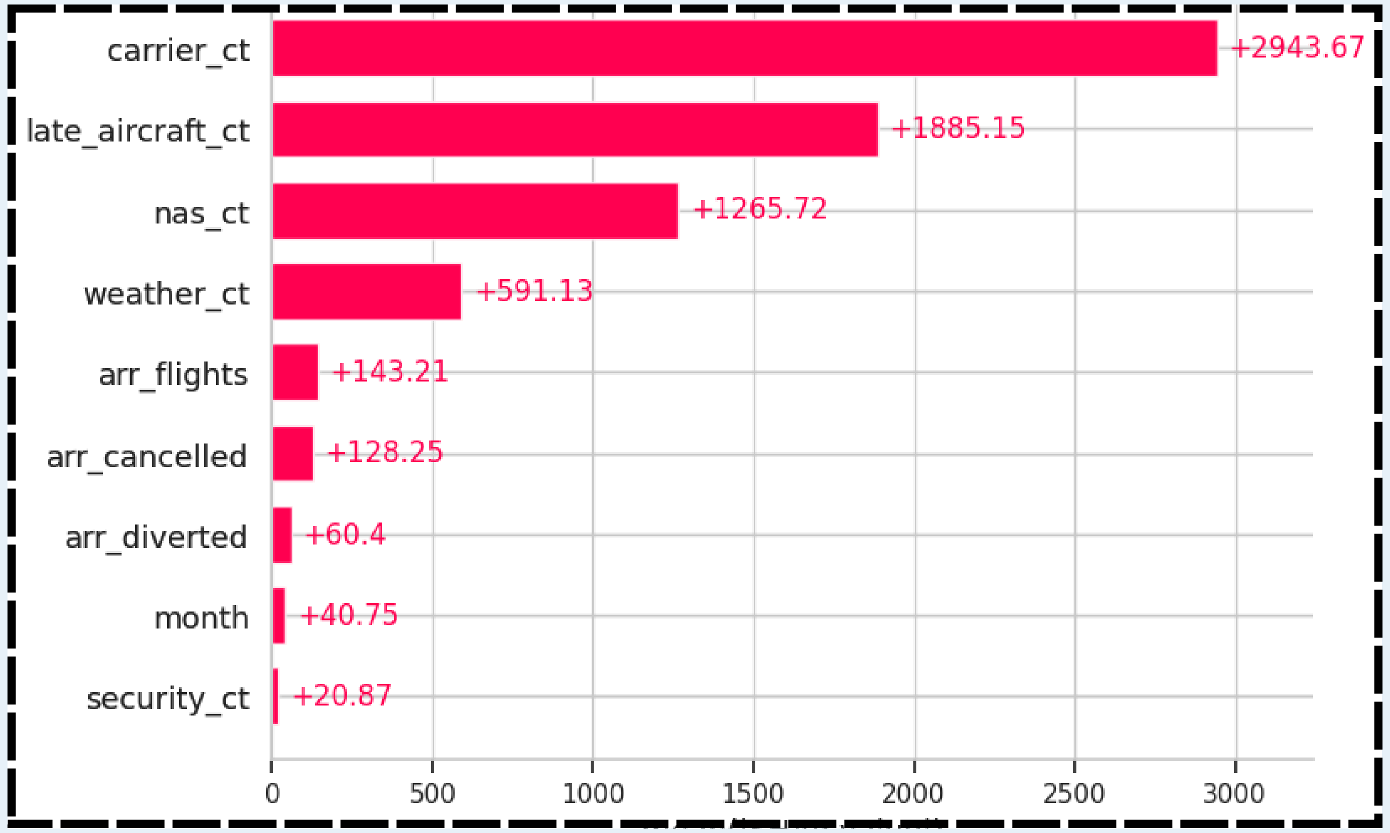
Why Random Forest?

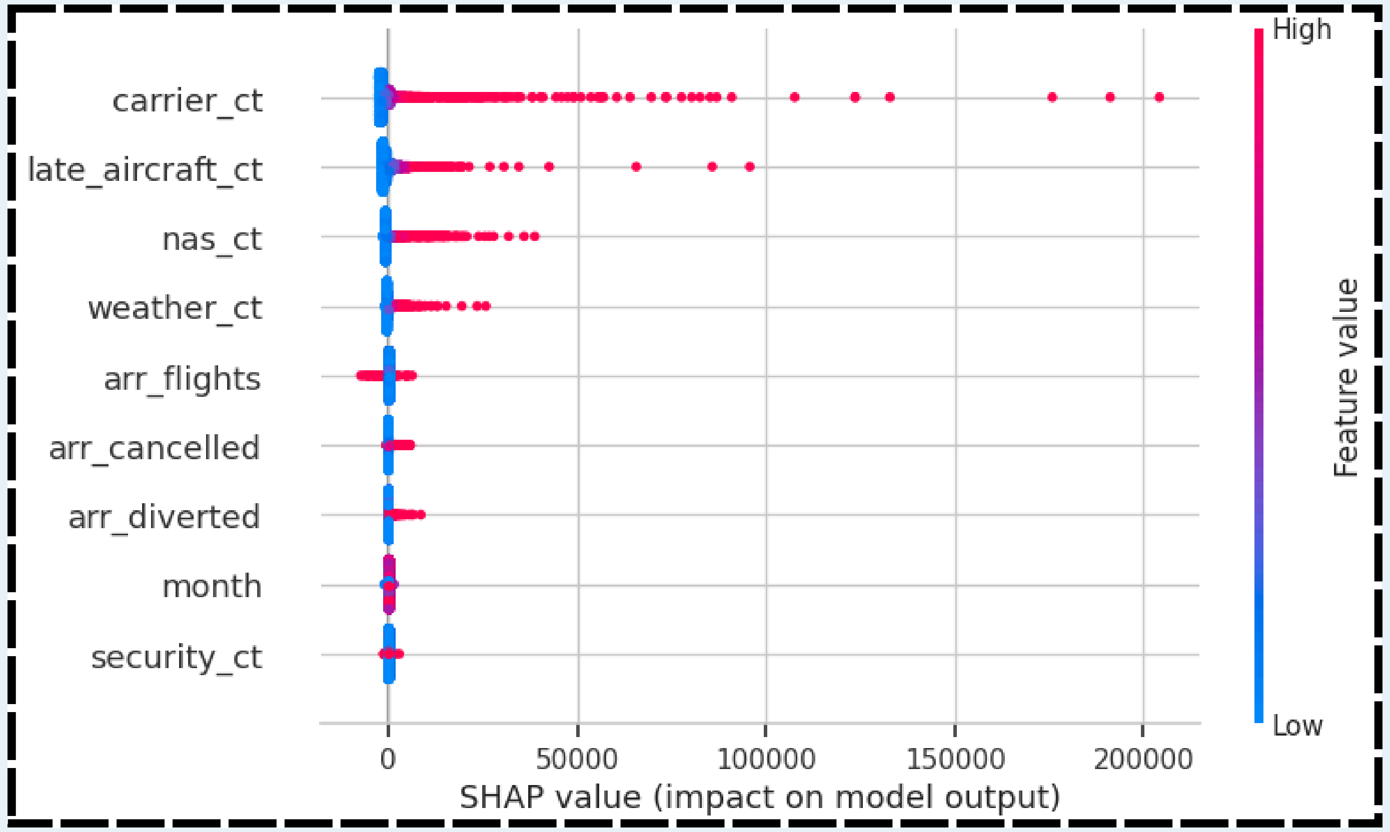
Chosen for its high accuracy and interpretability with SHAP. Handles mixed feature types and avoids overfitting with minimal tuning. Performs well on both classification and regression tasks for delays.

Society of Business

INTERPRETING & PRIORITIZING DELAY CAUSES

Using Explainable AI and Operational Focus





Model Explainability with SHAP

- carrier_ct and late_aircraft_ct are the most impactful features driving delay predictions
- SHAP assigns each prediction a set of values explaining how much each feature contributed
- High carrier_ct and late_aircraft_ct values = strong push toward predicted delay
- This enables transparent, airline-specific decisions — not black-box guessing

Operational Adjustability Index

- We designed a custom index to prioritize delays airlines can control
- OAI emphasizes carrier_ct and late_aircraft_ct over weather or NAS delays
- Flights with high SHAP and high OAI = top candidates for operational improvements

ACTIONABLE RECOMMENDATIONS FOR AIRLINES

From insights to operational impact

Actionable Recommendations

1. Shift High-Risk Flights to Earlier Time Slots
Most delays occur in the evening — reschedule vulnerable routes earlier

2. Improve Turnaround Time for Late Aircraft
One of the top SHAP + OAI features — invest in turnaround ops

3. Prioritize Flights with High OAI Scores
Focus on delays you can control: carrier & aircraft-related

4. Proactive Passenger Communication
Use model predictions to pre-alert passengers and reduce dissatisfaction

5. Consider SHAP-integrated Ops Dashboard
Embed SHAP insights into scheduling tools for real-time support

Impact Summary

- 99.8% delay prediction accuracy with full model transparency
- SHAP identifies top delay drivers
- OAI focuses airline attention on actionable causes
- Framework ready for integration into airline ops or dashboards

Next Steps / Deployment

- Integrate model into airline scheduling systems
- Automate daily delay risk reports using SHAP + OAI
- Run A/B tests on adjusted schedules for ROI tracking
- Expand model to include weather forecasts & live feeds

THANK YOU!