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

Sahil Jain (22118064)
IIT Roorkee

Problem Statement

- Flight delays lead to passenger dissatisfaction and increased operational cost ranging from fuel expenses and crew repositioning to missed connections and reputational damage.

- Understanding the underlying causes of these disruptions and proactively anticipating them is paramount for enhancing operational efficiency, improving customer satisfaction, and fostering a more reliable air travel ecosystem.

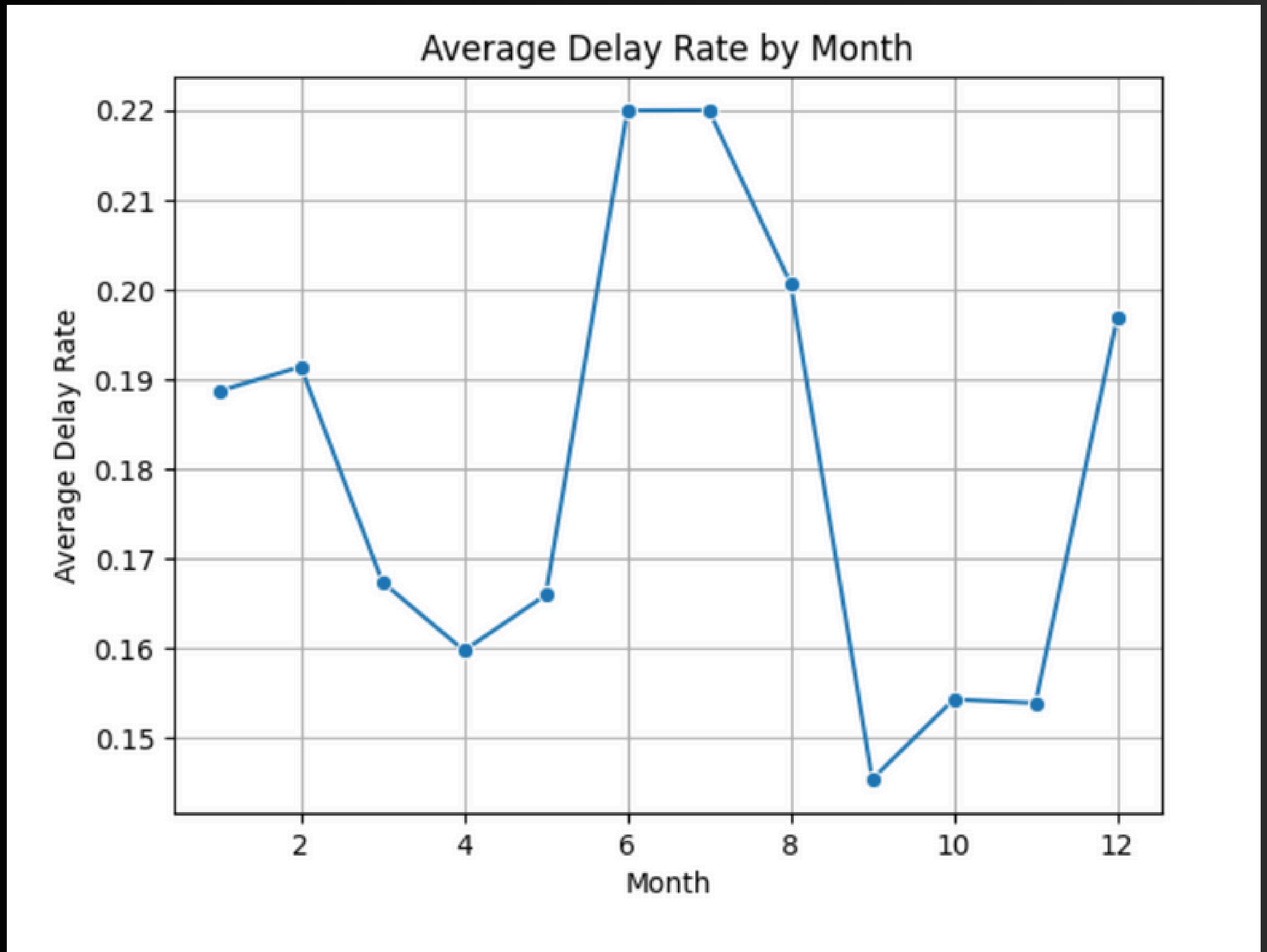
Project Objective

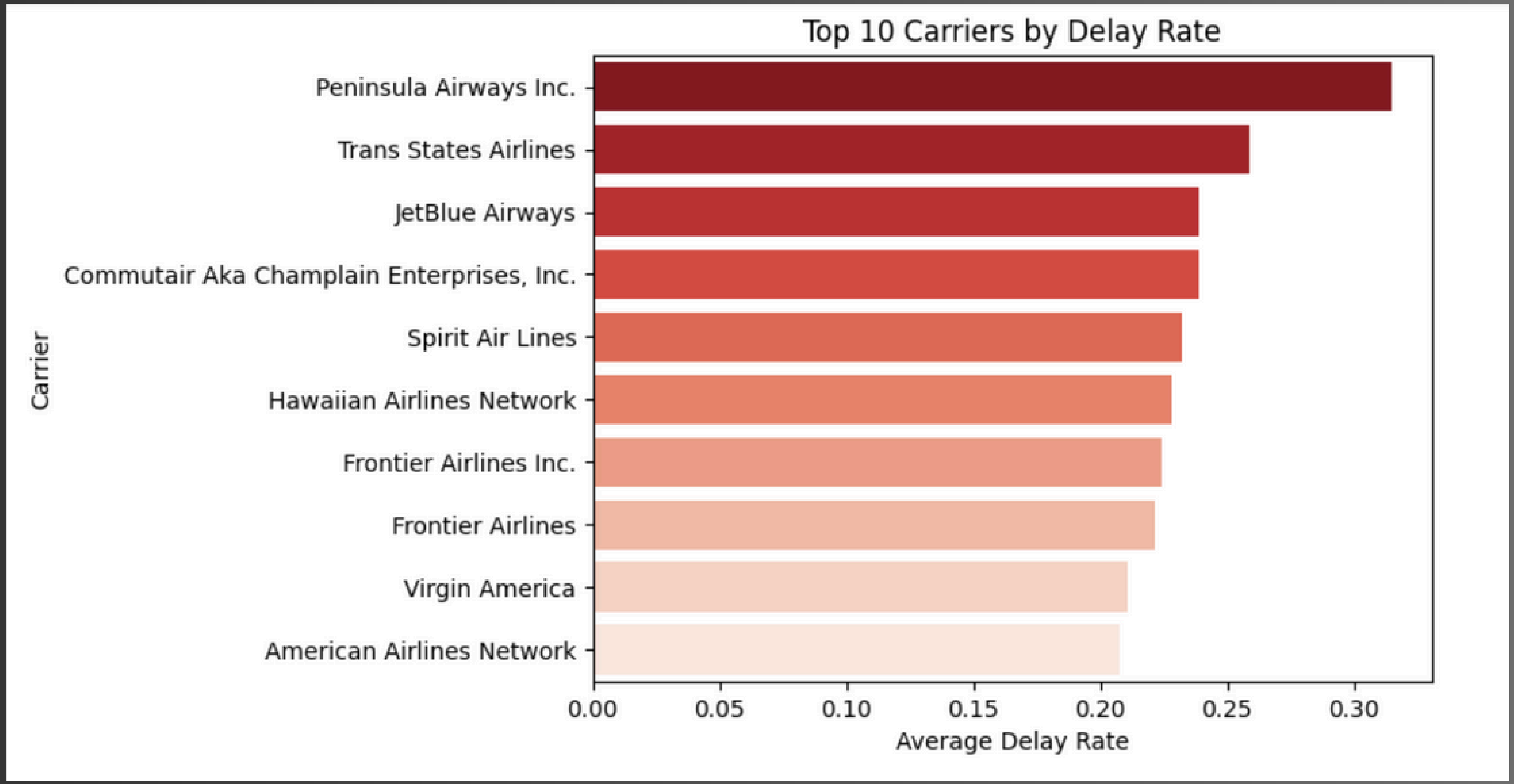
- Uncover Hidden Patterns: Conduct an in-depth exploratory data analysis (EDA) to identify recurring trends, influential factors, and significant correlations contributing to flight delays.

- Develop Predictive Capability: Build a robust analytical model capable of predicting the likelihood or duration of flight delays, providing an early warning system for stakeholders.

- Generate Actionable Insights: Formulate data-backed recommendations and strategic guidance for airlines and relevant stakeholders to mitigate delay occurrences and enhance operational resilience.

Exploratory Data Analysis

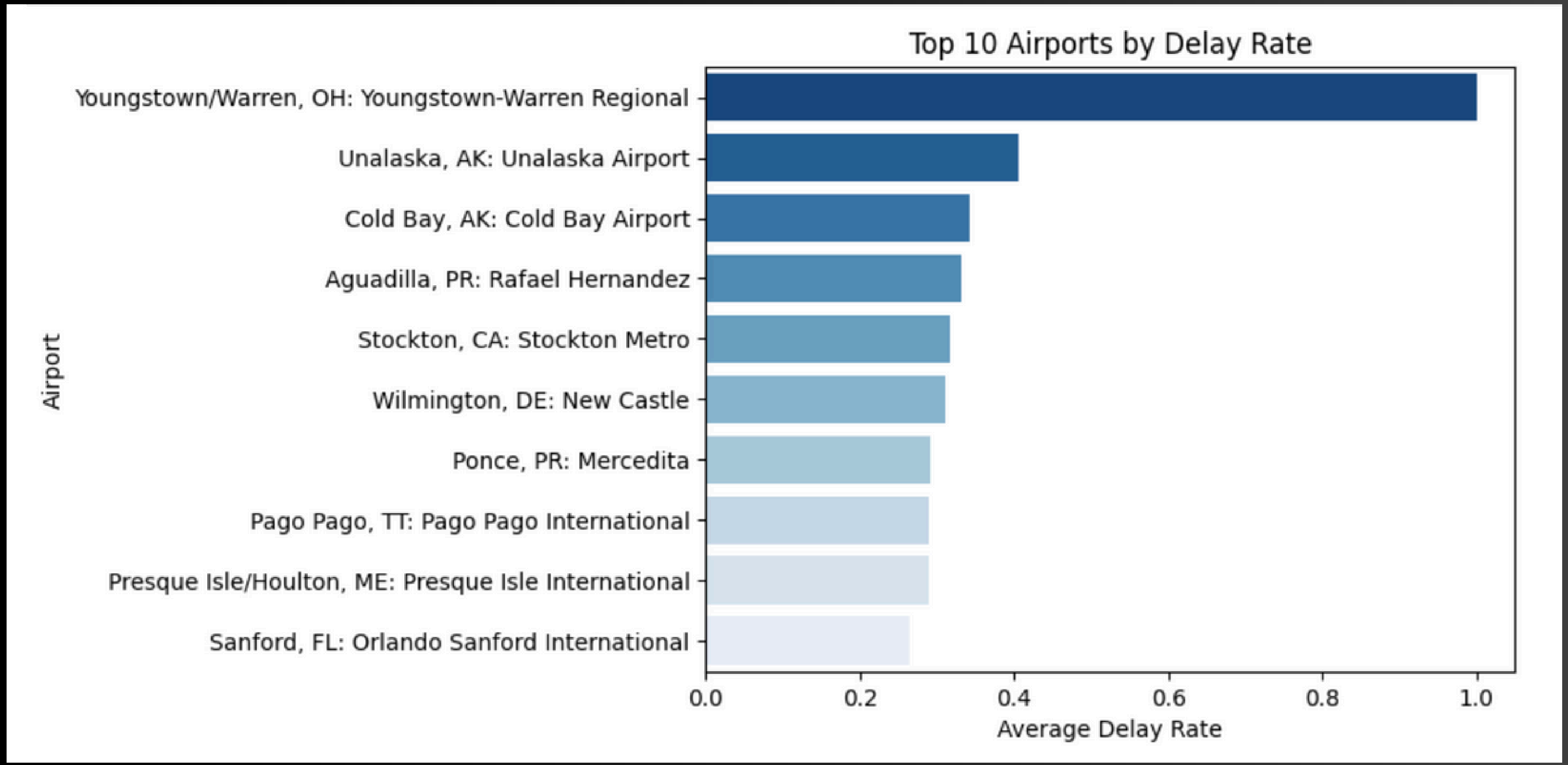


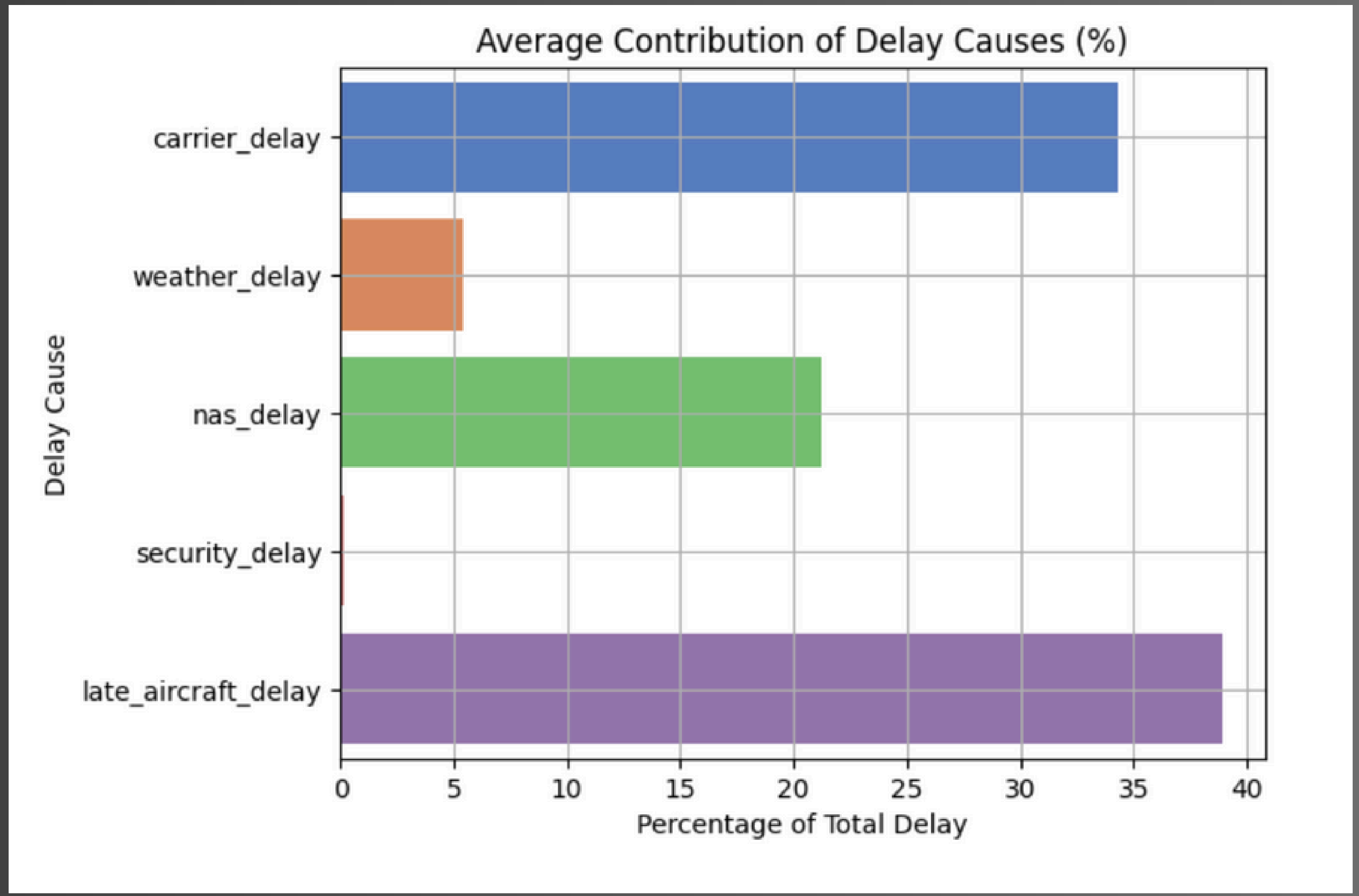


- Delay rates peak in June and July, suggesting seasonal congestion, likely due to summer travel.
- A sharp decline in September–November, which aligns with post-summer and pre-holiday lull in air traffic.
- Proactive planning and resource allocation during peak months can help mitigate delay surges.

- Peninsula Airways and Trans States Airlines have the highest average delay rates among all carriers.
- Majority of top delayed carriers are regional or low-cost airlines, often operating on tighter schedules.
- These carriers may benefit from better schedule buffer and resource management to improve reliability.

Exploratory Data Analysis





- Youngstown-Warren Regional airport has an exceptionally high delay rate, followed by Unalaska and Cold Bay.
- Smaller or remote airports dominate this list — likely due to limited infrastructure or weather exposure.
- Enhancing operational resilience and local coordination could reduce these high delay percentages

- Carrier delay and Late aircraft delay together contribute nearly 75% of total delays.
- Weather and NAS delays are present but secondary; security delays are negligible.
- Airlines should prioritize internal process optimization over external factors.

Root Cause Analysis

1. Delay Reason Breakdown

- Carrier Delay (avg ~30%): Stems from airline operations — e.g., crew issues, maintenance, paperwork. Controllable and should be prioritized.
- Late Aircraft Delay (avg ~45%): Most significant cause. It reflects cascading delays from earlier flights and tight aircraft turnaround times.
- NAS Delay (~15%): Caused by air traffic congestion or navigation system constraints. Partially controllable through better coordination.
- Weather Delay (~10%): Seasonal/weather-based — airports in snowy or storm-prone areas show spikes.
- Security Delay (<1%): Negligible overall impact.

Insight: Over 75% of delays are from causes within the airline's control — high potential for optimization.

2. Operational Bottlenecks Identified

- Tight turnaround scheduling: Delays from incoming flights (late aircraft) create a domino effect — major source of delay escalation.
- Under-resourced ground teams: Regional carriers with fewer backup teams see more carrier-related delays.
- Inadequate buffer during peak travel months: Summer and holiday seasons have highest delay rates due to over-scheduling.
- Congestion at key hubs: Airports like ATL, ORD show delays due to air traffic density and gate unavailability.

Insight: Focused interventions like schedule padding, spare aircraft placement, and ground staff scaling during peak times can significantly reduce delays.

Predictive Modeling and Performance Metrics

Objective:

Predicting – (1) if a flight will be delayed or not and (2) the delay duration

Models used:

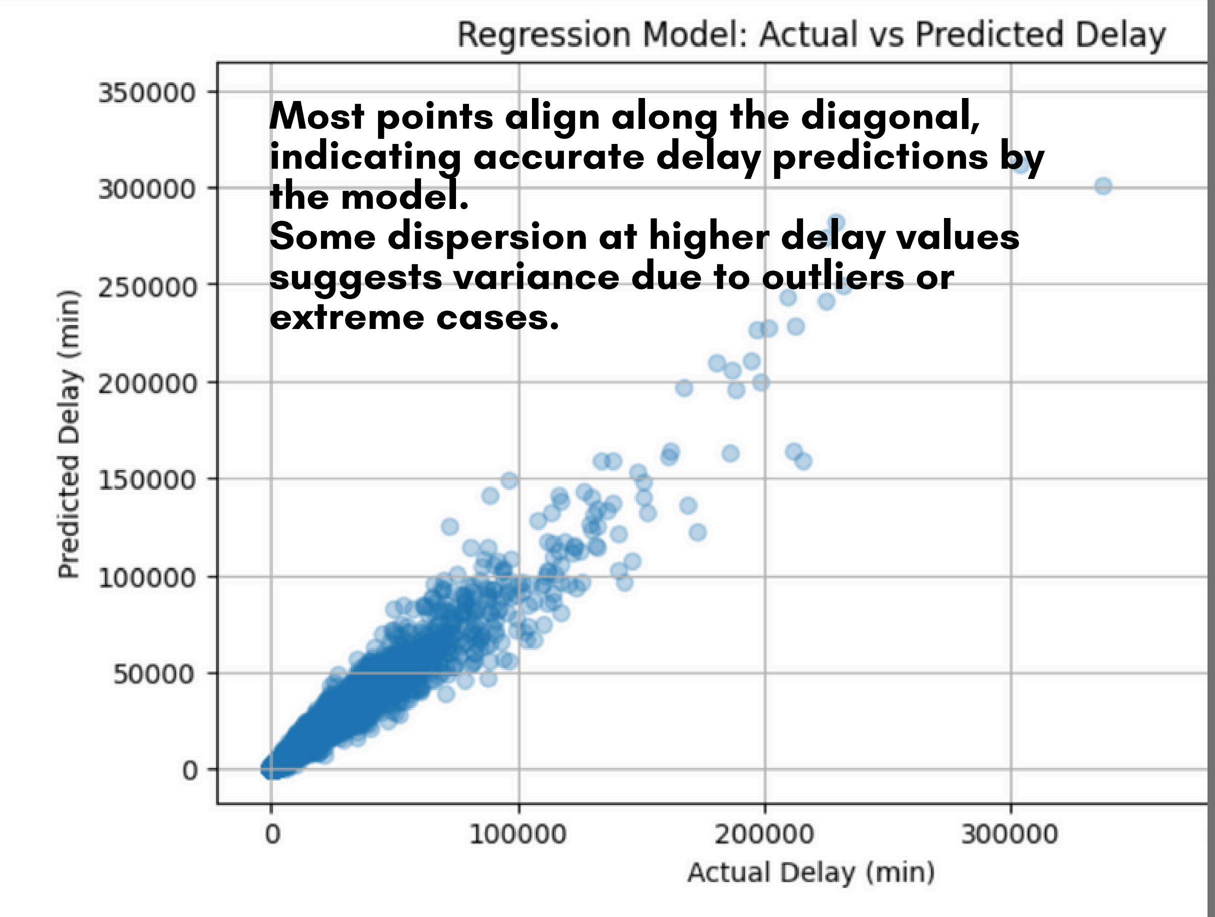
- Random Forest Classifier (for Yes/No delay)
- Random Forest Regressor (for delay duration)

Evaluation Metrics:

- Classification: Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC Curve
- Regression: MAE, RMSE

[66]:

	Actual Delay (min)	Predicted Delay (min)
0	304.0	302.580000
1	3294.0	3989.810000
2	304.0	422.670000
3	237.0	173.895000
4	104.0	124.037833
5	1937.0	2244.240000
6	642.0	751.670000
7	170.0	407.030000
8	5120.0	3845.680000
9	254.0	118.472667



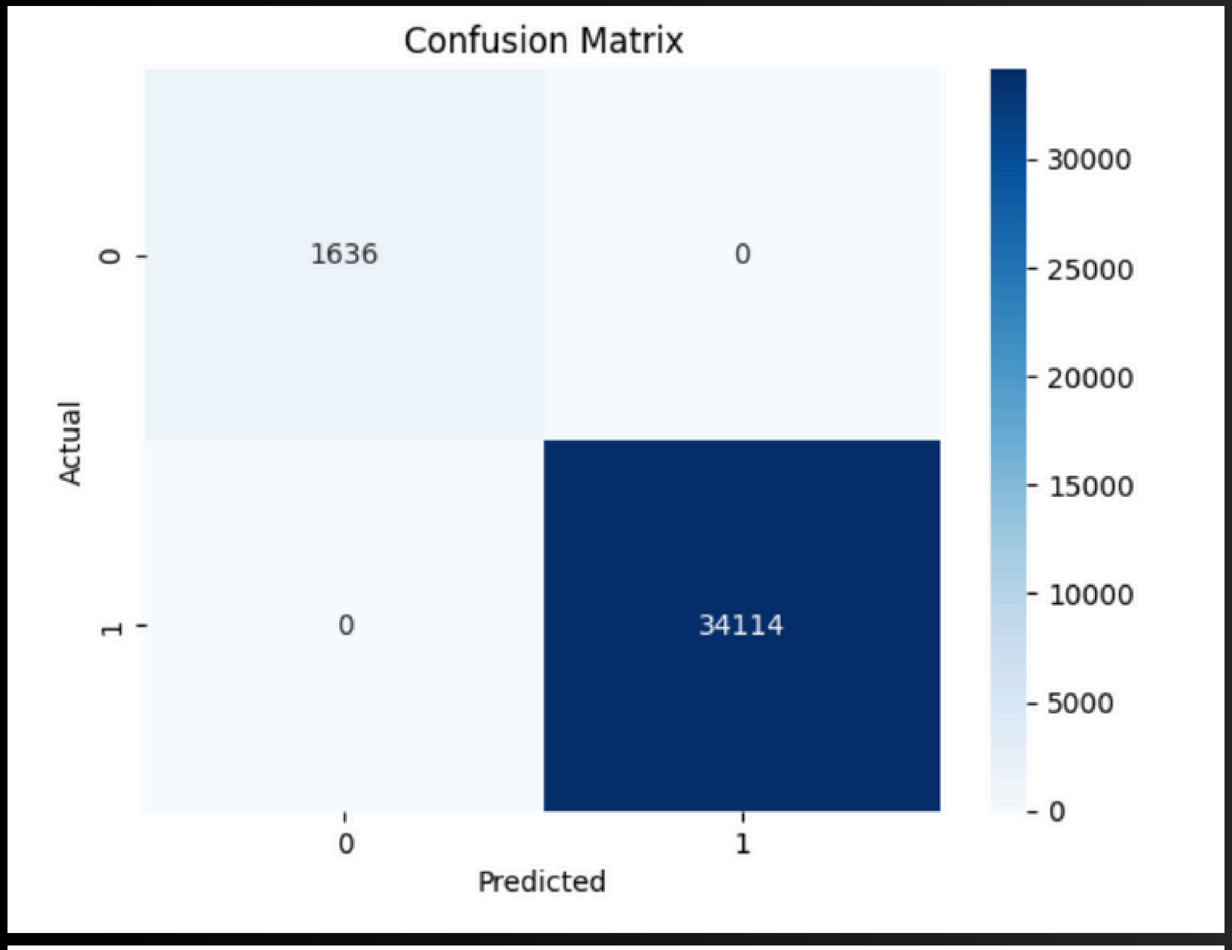
[19]:

```
## Metrics
print("Classification Metrics:")
print("Accuracy:", accuracy_score(y_test_class, y_pred_class))
print("Precision:", precision_score(y_test_class, y_pred_class))
print("Recall:", recall_score(y_test_class, y_pred_class))
print("F1 Score:", f1_score(y_test_class, y_pred_class))
print("ROC AUC Score:", roc_auc_score(y_test_class, y_proba_class))
```

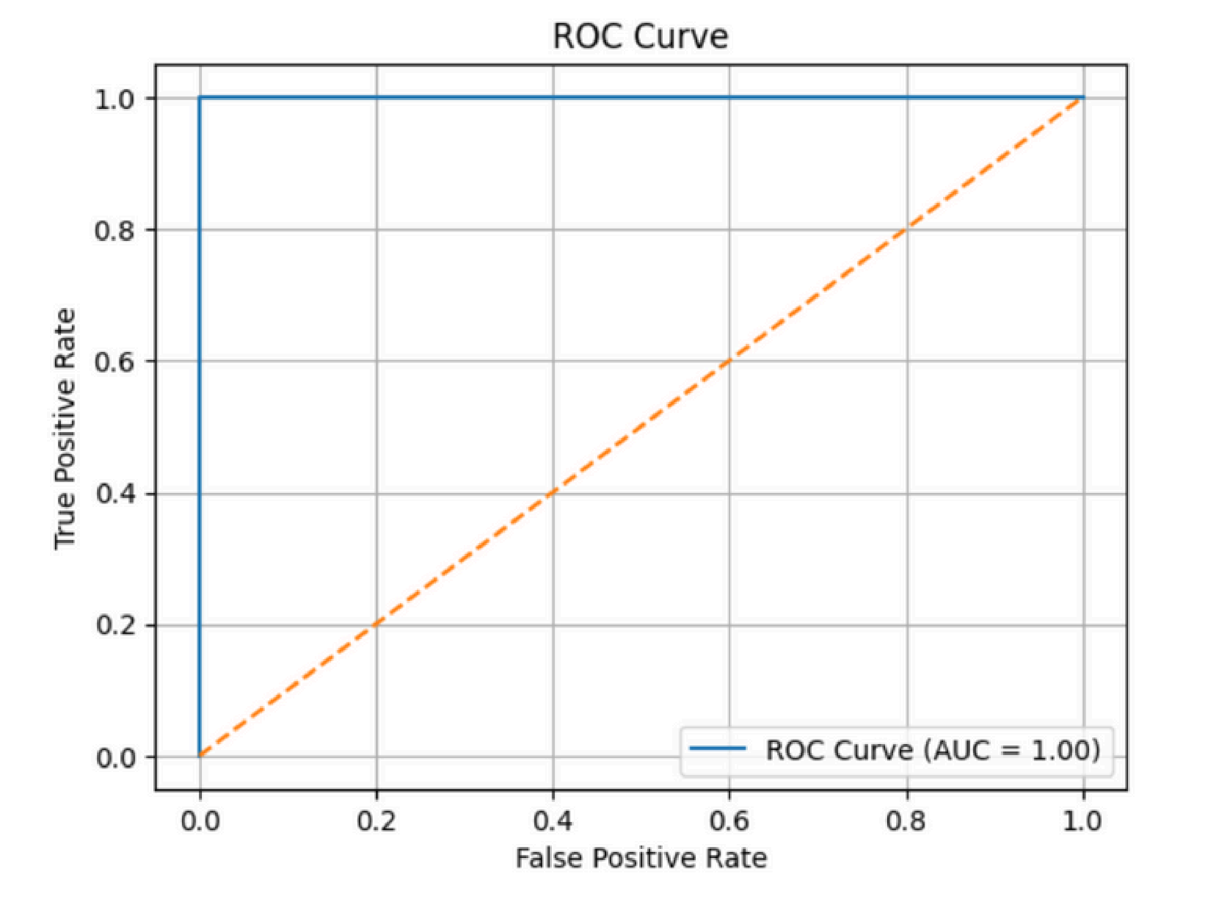
```
Classification Metrics:
Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1 Score: 1.0
ROC AUC Score: 1.0
```

- Perfect classification performance with 100% scores across Accuracy, Precision, Recall, F1, and ROC AUC, indicating the model correctly predicted all flight delay instances in the test set.
- Highlights potential data leakage or imbalance, as such perfect metrics are extremely rare in real-world scenarios and warrant deeper validation of data quality and model generalization.

Predictive Modeling and Performance Metrics



- The model achieved perfect classification, with all 1636 negative samples (class 0) and all 34,114 positive samples (class 1) correctly identified — resulting in no false positives or false negatives.
- The matrix shows complete separation, with the top-left value (1636) being the true negatives and the bottom-right value (34114) being the true positives, confirming a 100% accuracy on the test set.



- The ROC curve reaches the top-left corner, indicating that the model achieves maximum true positive rate with minimal false positives across all thresholds.
- AUC = 1.0 indicates the model perfectly separates delayed from non-delayed flights.

Regression Model (Delay Duration)

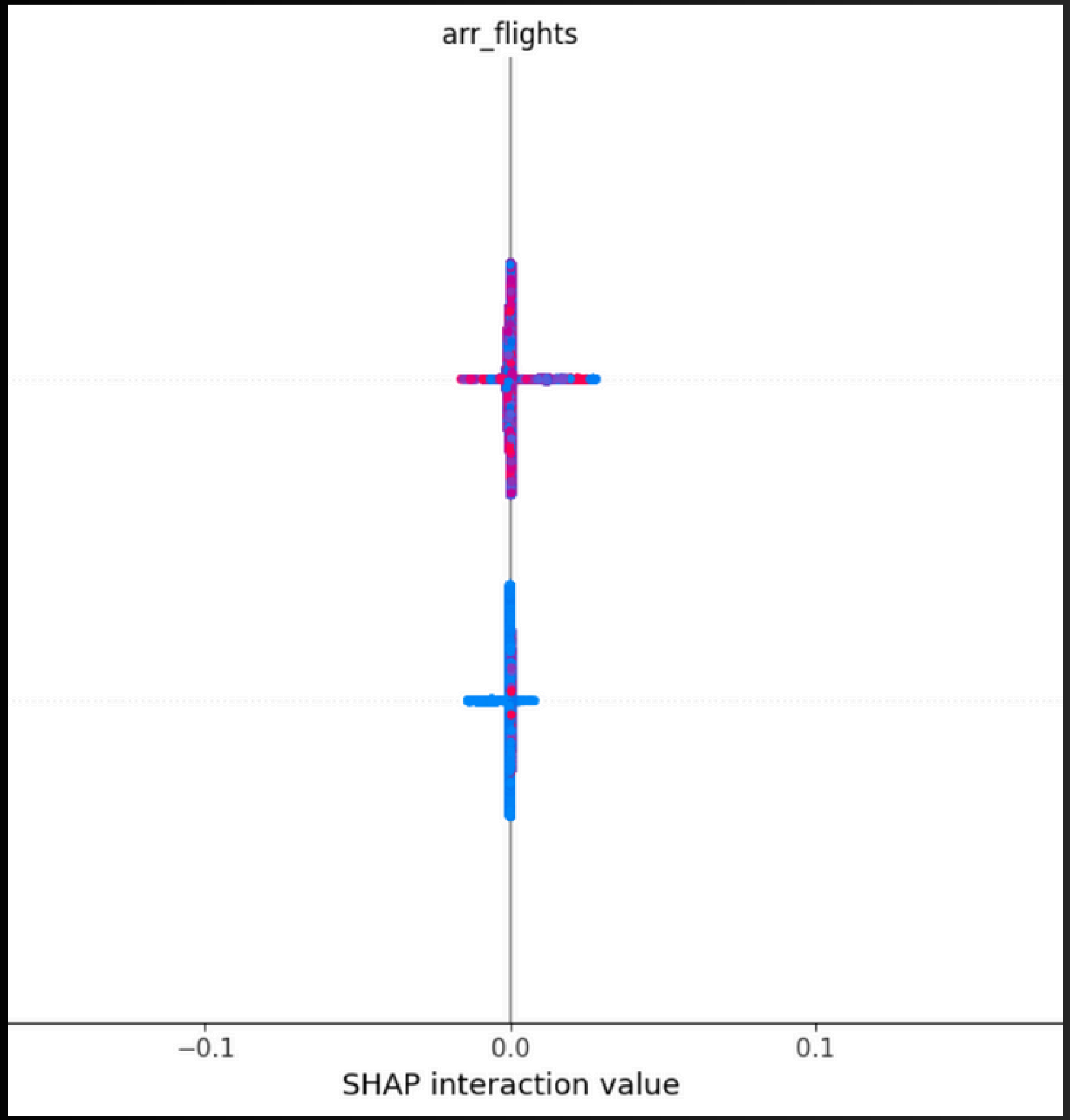
```python
reg = RandomForestRegressor(random_state=42)
reg.fit(X_train, y_train_reg)
y_pred_reg = reg.predict(X_test)

print("Regression Metrics:")
print("MAE:", mean_absolute_error(y_test_reg, y_pred_reg))
print("RMSE:", np.sqrt(mean_squared_error(y_test_reg, y_pred_reg)))
```

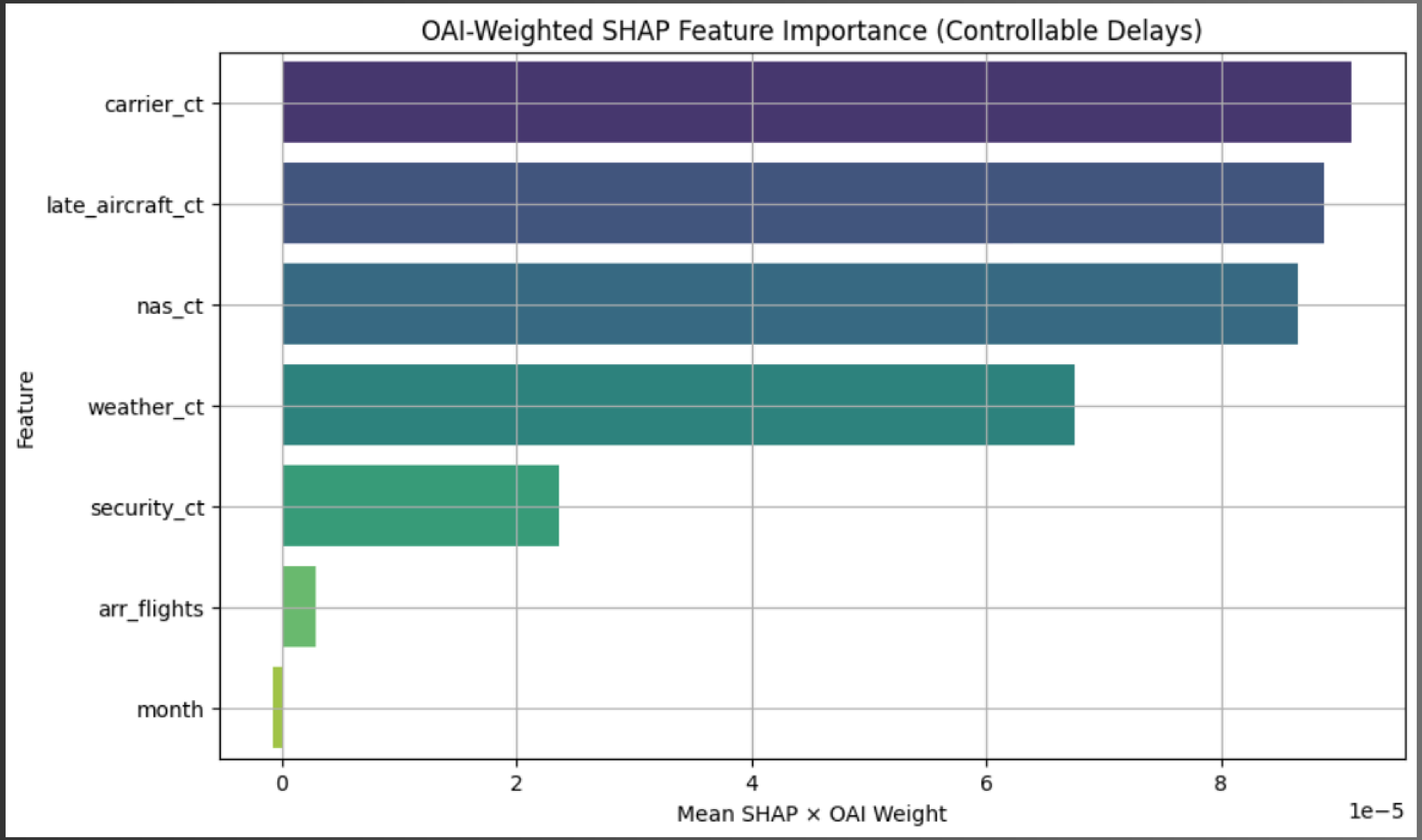
```
Regression Metrics:
MAE: 687.3640357549917
RMSE: 2271.5855834654303
```

- The model predicts delay durations with an average error of around 11.5 hours (MAE) — accurate for most regular delays.
- A high RMSE (around 38 hours) suggests a few extreme outliers where the model under- or over-predicted by large margins.

SHAP Explainability & Operational Adjustability Index

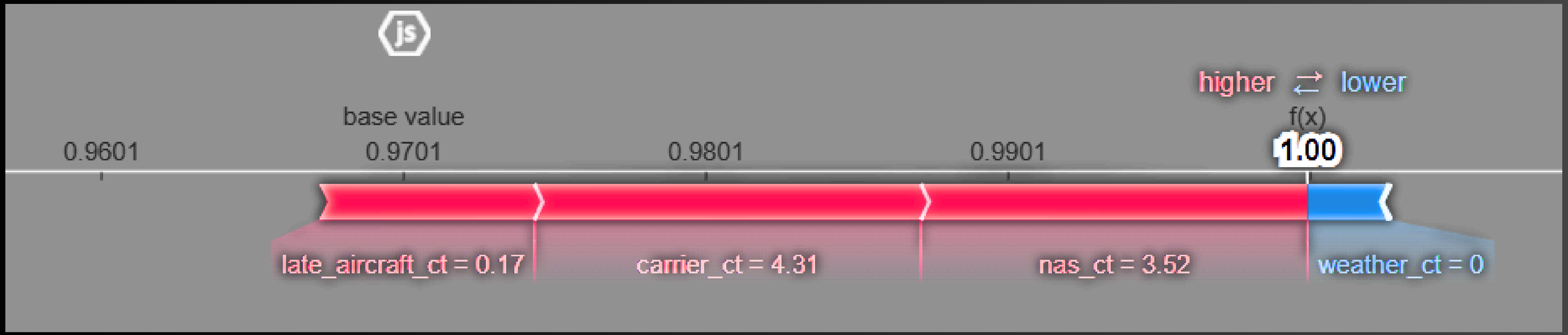


- OAI: Custom metric combining carrier_delay and late_aircraft_delay to focus on controllable delays



- arr_flights, carrier_ct, and late_aircraft_ct are the most influential features in predicting whether a flight is delayed.
- SHAP reveals how each feature positively or negatively contributes to delay probability across all predictions.
- After applying Operational Adjustability Index (OAI) weights, carrier_ct and late_aircraft_ct become dominant — highlighting controllable delay factors.
- This plot helps airlines prioritize internal improvements (like aircraft readiness, gate scheduling) over external issues (like weather).

Prediction Breakdown & Actionable Recommendations



- This force plot explains how each feature pushed the prediction toward a delayed or on-time outcome for a single flight.
- Features in red (e.g., late_aircraft_ct, carrier_ct) increased the likelihood of delay, while blue features (e.g., weather_ct) pushed toward an on-time prediction.
- The width of each bar reflects the magnitude of that feature's influence on the model's decision.

Insight:
This level of interpretability allows airlines to understand why a specific flight was flagged as high-risk, and take targeted action.

Strategic Recommendations for Delay Mitigation:

- Add turnaround and gate buffer time during peak months (June–August, Dec) to reduce cascading delays from late aircraft.
- Optimize staffing and resource deployment at delay-prone airports — especially for baggage handling, fueling, and crew readiness.
- Use delay predictions to send early alerts to passengers and reassign crew resources in real time.
- Prioritize efforts on carrier and late aircraft delays, as identified by the SHAP + OAI model, since they are both high-impact and controllable.

Thank you!