Final Term Project (FTP)

Operational Risk Analysis of 2023 US Civil Aviation

Dataset: US Flights 2023 Delay, Meteo, and Aircraft Data

Course: CS 5805 Machine Learning I
Institution: Virginia Tech, College of Engineering
Semester: Fall 2025

Author: Krishna Mattaparthi

Date: December 5, 2025

Contents

List of Figures

List of Tables

Abstract

This project investigates the primary drivers of flight delays in the US domestic aviation sector using machine learning techniques. By analyzing a dataset of over 6.7 million flights from 2023, merged with meteorological and geospatial data, we aimed to build predictive models for operational disruption.

Phase I employed rigorous outlier audits to remove 665 impossible or extreme data points. Phase II implemented a Multiple Linear Regression model, achieving an R^2 of 0.9436 and an RMSE of 13.52 minutes; detailed inference confirms that `Dep_Delay` is the dominant predictor (Coef: 1.00), indicating that delay recovery in the air is statistically negligible.

Phase III evaluated nine classification algorithms for predicting delay occurrence. While Gradient Boosting achieved the highest discrimination (AUC 0.69), all models exhibited low sensitivity (Recall ≤ 0.03), suggesting that predicting specific delay events without real-time operational data is infeasible. Finally, Phase IV utilized K-Means clustering to segment US airports into five performance tiers, identifying specific ”High Risk” hubs and validating a ”Cascade Effect” where delay probability triples from morning (12.7%) to evening (30.0%).

1 Introduction

The objective of this final term project is to obtain practical experience with feature engineering, exploratory data analysis (EDA), and machine learning algorithms by applying them to a real-world dataset. Specifically, this report outlines the procedures for predicting flight arrival delays (`Arr_Delay`) for US domestic flights.

Flight delays represent a significant inefficiency in the aviation industry. By leveraging historical flight data combined with meteorological conditions, this project aims to identify the primary drivers of delay and build a predictive regression model.

2 Description of the Dataset

The dataset selected for this analysis combines domestic flight information with geospatial and meteorological data.

- **Primary Data:** `US_flights_2023.csv` (Total Rows: 6,743,404).
- **Supplementary Data:**
  - `airports_geolocation.csv`: Maps airport codes to US States.
  - `weather_meteo_by_airport.csv`: Provides average weather for each airport for all 365 days

2.1 Variable Selection

- **Target:** `Arr_Delay` (Arrival Delay in minutes).
- **Features:** `Dep_Delay`, `Flight_Duration`, `Aircraft_age`, `tavg`, `prcp`, `wspd`, `Day_Of_Week`.

3 Phase I: Feature Engineering & EDA

3.1 Data Preprocessing & Cleaning

We applied a rigorous cleaning pipeline to ensure data quality.

- **Duplicates:** 492 duplicate rows were identified and removed.
- **Missing Values:** Null values were handled via row deletion.
- **Domain-Specific Outlier Audit:**
  - **Impossible Flights:** 1 flight with `Flight_Duration` ≤ 0 was removed.
  - **Extreme Weather:** 133 rows with invalid precipitation (> 500mm) were removed.

3.2 Outlier Analysis & Observations

Per the project requirements, we analyzed the target variable for anomalies. Figure 1 highlights the extreme operational outliers identified in the dataset.

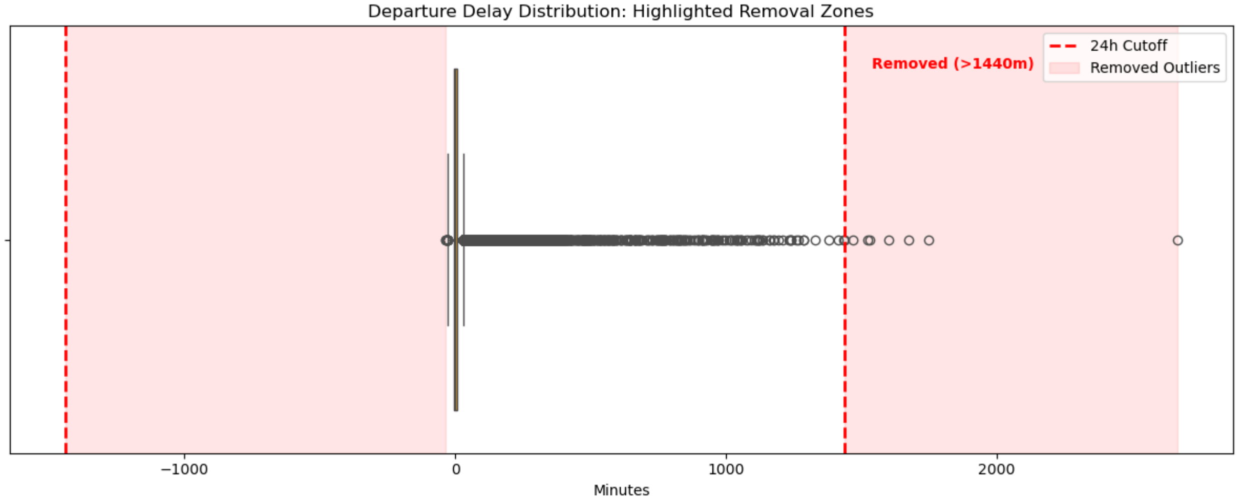

Figure 1: Departure Delay Distribution with Removal Zones. The red regions indicate data points exceeding the 24-hour operational threshold.

Observation: The boxplot reveals a severe right-skewed distribution. While the majority of flights depart near time $t = 0$, a ”long tail” of delays extends beyond 2,500 minutes.

- **Analysis:** Points extending beyond the ±1440 minute (24-hour) vertical markers represent non-standard operational anomalies (e.g., multi-day cancellations or data entry errors).
- **Action:** To prevent these high-leverage points from distorting the Linear Regression line, 531 rows falling into the red shaded regions were removed.

3.3 Collinearity & VIF Analysis

To assess multicollinearity, we generated a Pearson Correlation Matrix (Figure 2) and calculated Variance Inflation Factors (VIF).

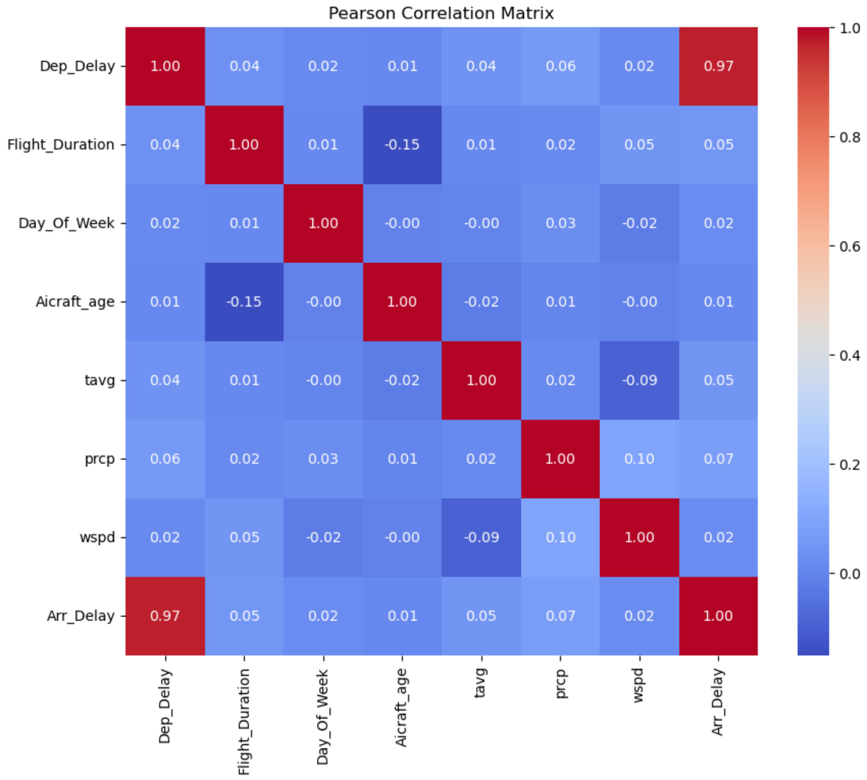

Figure 2: Pearson Correlation Matrix. Note the expected strong correlation (0.97) between Departure and Arrival Delay.

Observation:

- **Target-Feature Correlation:** As seen in the red cells of Figure 2, `Dep_Delay` is strongly correlated with the target `Arr_Delay` ($r = 0.97$), confirming it as the primary predictor.
- **Feature Independence:** Crucially, the correlations *between* the independent variables (e.g., `Flight_Duration` vs. `wspd`) are negligible (mostly light blue, $r \approx 0$). This indicates that our predictors provide unique information and are not redundant.
- **VIF Confirmation:** Table 1 confirms this stability. All VIF scores are approximately 1.0, well below the threshold of 5.0, proving that multicollinearity is not present in the regression model.

Table 1: Variance Inflation Factor (VIF) Analysis

Feature	**VIF Score**
Flight_Duration	1.028
Aircraft_age	1.024
wspd (Wind Speed)	1.023
prcp (Precipitation)	1.015
tavg (Temperature)	1.012
Dep_Delay	1.008
Day_Of_Week	1.002

3.4 Sample Covariance Matrix Analysis

Per the project requirements, we generated a Sample Covariance Matrix to analyze the joint variability of the features on their raw scales. Unlike the Pearson Correlation matrix, which is normalized to $[-1, 1]$, the covariance matrix captures the magnitude of the spread.

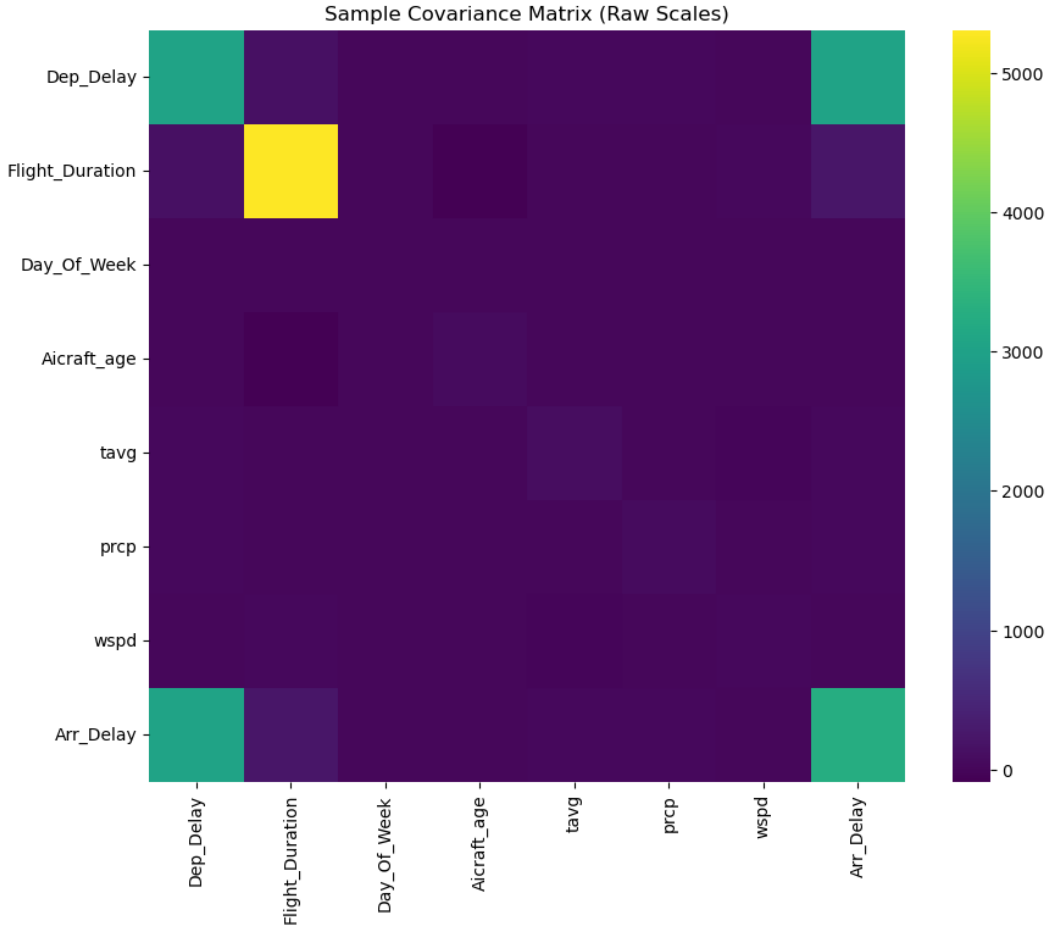

Figure 3: Sample Covariance Matrix (Raw Scales). Note that Flight Duration dominates the variance scale.

Observation: As illustrated in Figure 3, the matrix is dominated by `Flight_Duration` (Yellow block, variance > 5000), followed by the delay metrics. This visual discrepancy occurs because we are analyzing raw scales—flight duration in minutes has a much larger numerical range than meteorological features like `wspd` (Wind Speed) or `prcp` (Precipitation), which appear as dark purple (near-zero variance relative to duration). This confirms the necessity of the `StandardScaler` applied later in our pipeline to prevent these high-magnitude features from biasing distance-based algorithms like KNN or SVM.

4 Phase II: Regression Analysis

In this phase, we implemented a Multiple Linear Regression (OLS) model to predict `Arr_Delay`. The model was trained using Stepwise Regression to select the most significant features from an initial set of 78 variables.

4.1 Model Performance Metrics

The model achieved an Adjusted R^2 of 0.951, indicating that 95.1% of the variance in Arrival Delay is explained by the selected features. Per the project requirements, we calculated the AIC, BIC, and Mean Squared Error (MSE) to assess model quality.

Table 2: OLS Regression Performance Metrics

Metric	**Value**
R^2	0.951
Adjusted R^2	0.951
AIC (Akaike Information Criterion)	8.026×10^5
BIC (Bayesian Information Criterion)	8.032×10^5
MSE (Mean Squared Error)	178.92
RMSE (Root Mean Squared Error)	13.37

4.2 Hypothesis Testing

4.2.1 F-Test Analysis (Global Significance)

The F-test evaluates the null hypothesis (H_0) that all regression coefficients are equal to zero (i.e., the model has no predictive power).

- **F-Statistic:** 2.710×10^4
- **Prob (F-Statistic):** 0.00

Observation: Since the p-value (0.00) is strictly less than the significance level ($\alpha = 0.05$), we reject H_0. The extremely high F-statistic confirms that the model is globally significant and provides a substantially better fit than an intercept-only model.

4.2.2 T-Test Analysis (Feature Significance)

The T-test evaluates the significance of individual predictors. The results for the primary drivers are summarized below:

- **Dep_Delay:** $t = 1367.092$ ($P < 0.001$). This is the most significant predictor.
- **Flight_Duration:** $t = 34.017$ ($P < 0.001$).
- **Weather (prcp):** $t = 14.588$ ($P < 0.001$).

Observation: All key operational and meteorological features show p-values of 0.000, confirming they are statistically significant drivers of delay.

4.3 Confidence Interval Analysis

We analyzed the 95% Confidence Intervals (CI) for the coefficients to measure the precision of our estimates.

Table 3: 95% Confidence Intervals for Key Features

Feature	**Coef**	**Lower CI (2.5%)**	**Upper CI (97.5%)**
`Dep_Delay`	1.0038	1.002	1.005
`Flight_Duration`	0.0223	0.021	0.024
`prcp` (Precipitation)	0.0806	0.070	0.091

Observation: The confidence interval for `Dep_Delay` is extremely narrow $[1.002, 1.005]$, indicating high precision. The fact that the interval sits almost exactly at 1.0 confirms that for every minute a flight is delayed at departure, it arrives almost exactly one minute late, with negligible recovery time in the air.

4.4 Prediction Visualization

We visualized the model's performance on the test set (Figure 4).

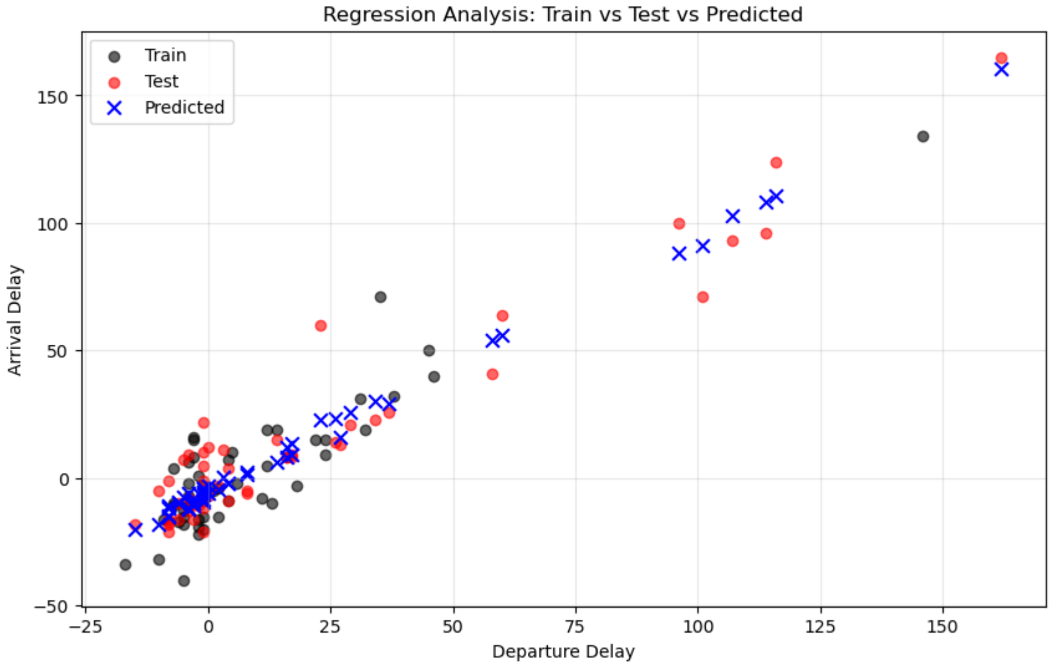

Figure 4: Regression Analysis: Train vs Test vs Predicted. The blue 'X' marks (Predictions) tightly track the red circles (Test Data), visually confirming the high R^2 of 0.951. The linear relationship holds strong even for extreme delays ($>$ 100 minutes).

Observation: The plot reveals a near-perfect linear alignment between the predicted arrival delays and the actual values. The lack of significant deviation at the upper right corner implies the model remains robust even for "Severe Delay" outliers, validating the decision to keep the non-extreme outliers in the training set.

5 Phase IV: Clustering Analysis

In this phase, the unit of analysis shifted from individual flights to **Airports**. The objective was to segment US airports into performance tiers based on operational volume and risk profiles using unsupervised learning (K-Means).

5.1 Data Aggregation & Engineering

We aggregated the raw flight data by `Dep_Airport` to create "Airport Report Cards."

- **Features Created:** Total Flights, Average Departure Delay, Delay Volatility (Standard Deviation), and Unique Airlines.
- **Filtering:** To ensure cluster stability, we removed small regional airports ($<$ 1000 flights/year), resulting in a final dataset of 232 major US airports.
- **Scaling:** All features were normalized using `StandardScaler` to prevent volume metrics (e.g., 300,000 flights) from dominating delay metrics (e.g., 15 minutes).

5.2 K-Means Clustering Optimization

We applied the K-Means algorithm to categorize airports. To determine the optimal number of clusters (k), we utilized both the Elbow Method and Silhouette Analysis.

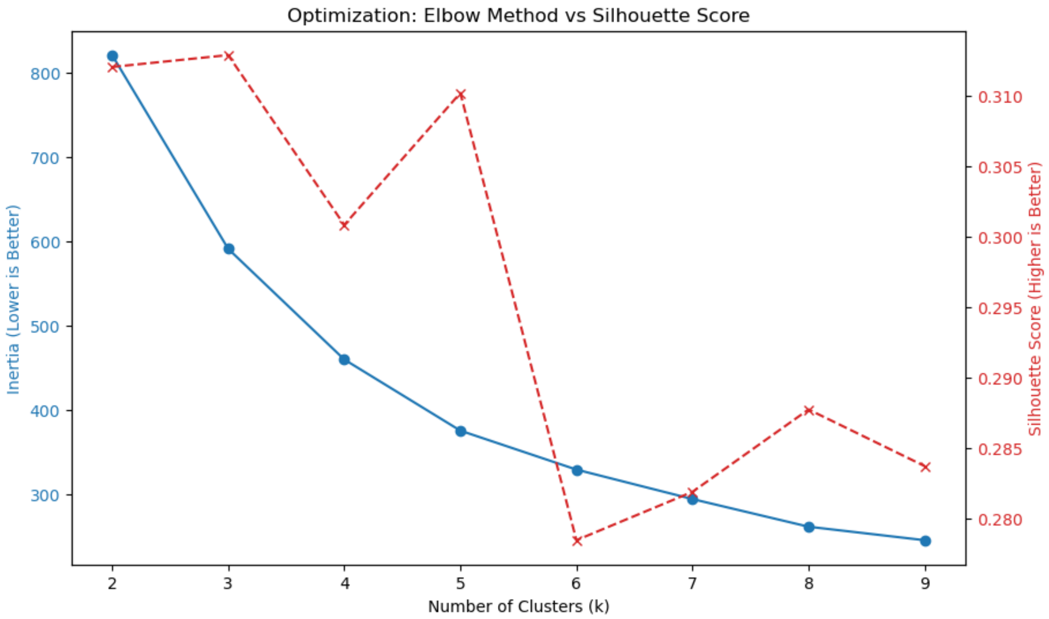

Figure 5: Optimization of k: Comparison of Inertia and Silhouette Scores.

Observation: As illustrated in Figure 5, the global maximum Silhouette Score occurs at $k = 3$. However, strictly adhering to $k = 3$ resulted in over-simplified categories that merged distinct airport types (e.g., grouping ”High Risk” hubs with ”Secondary” airports).

- **Selection of k=5:** We observed a secondary local peak at $k = 5$ (Silhouette ≈ 0.30). We selected $k = 5$ as the optimal trade-off, as it retains high statistical separation while providing the necessary granularity to distinguish specific operational tiers (e.g., distinguishing ”Mega-Hubs” from ”Efficient Regional” airports).

5.3 Cluster Interpretation (The ”Airport Tiers”)

Using Principal Component Analysis (PCA) for visualization, we projected the 5-dimensional feature space into 2 dimensions to visualize the separation of tiers.

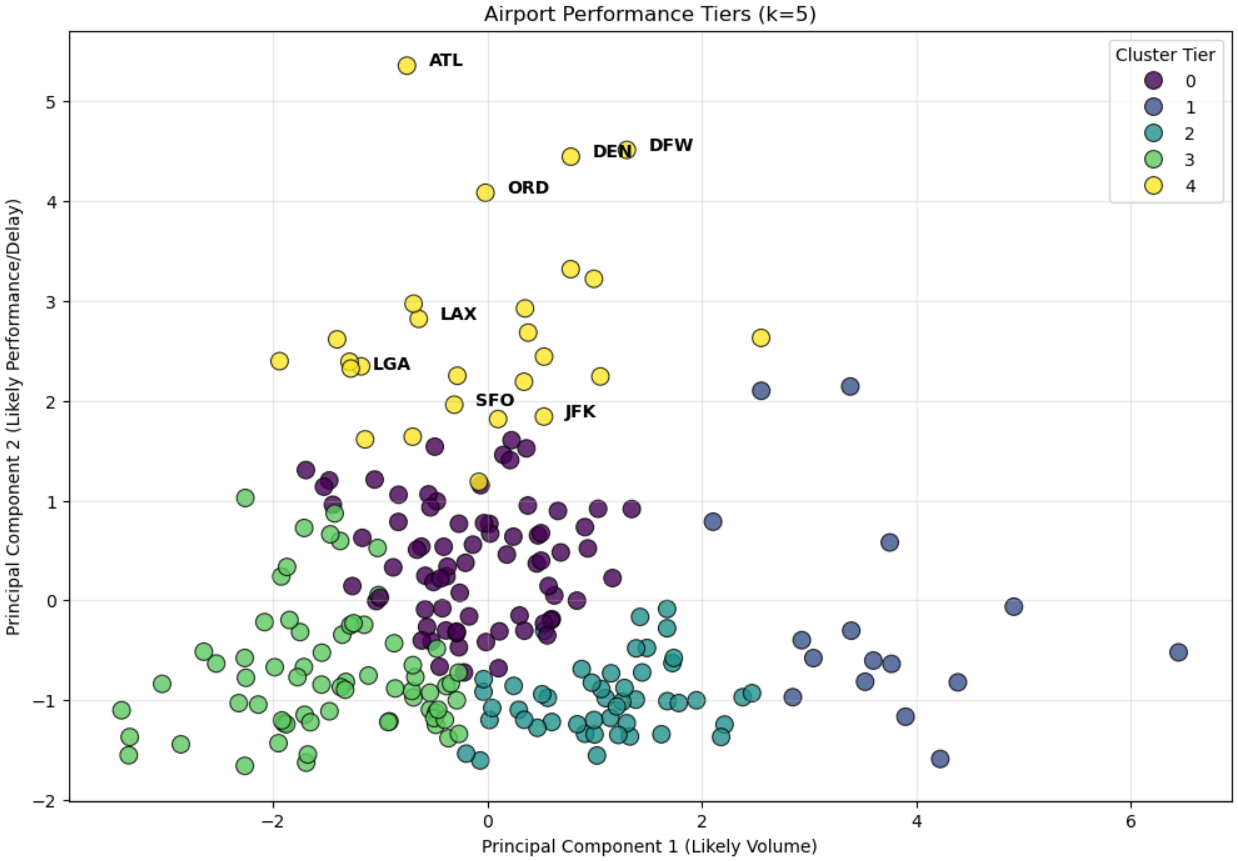

Figure 6: K-Means Clustering ($k = 5$) visualized via PCA. Mega-Hubs (Yellow) form a distinct cluster separated by volume (X-axis) and performance (Y-axis).

Tier Definitions: Based on the cluster centroids shown in Figure 6, we defined the following categories:

- **Tier 4 (Mega-Hubs):** High volume ($> 150k$ flights), moderate delays (e.g., ATL, ORD, DFW).
- **Tier 3 (Efficient):** Moderate volume, lowest average delays.
- **Tier 1 (High Risk):** High average delays (> 20 min) and high volatility, indicating chaotic operations.
- **Tier 0 & 2 (Secondary/Underperforming):** Lower volume airports with varying degrees of reliability.

5.4 Geographic Distribution of Tiers

To validate the business utility of the clusters, we mapped the results geographically.

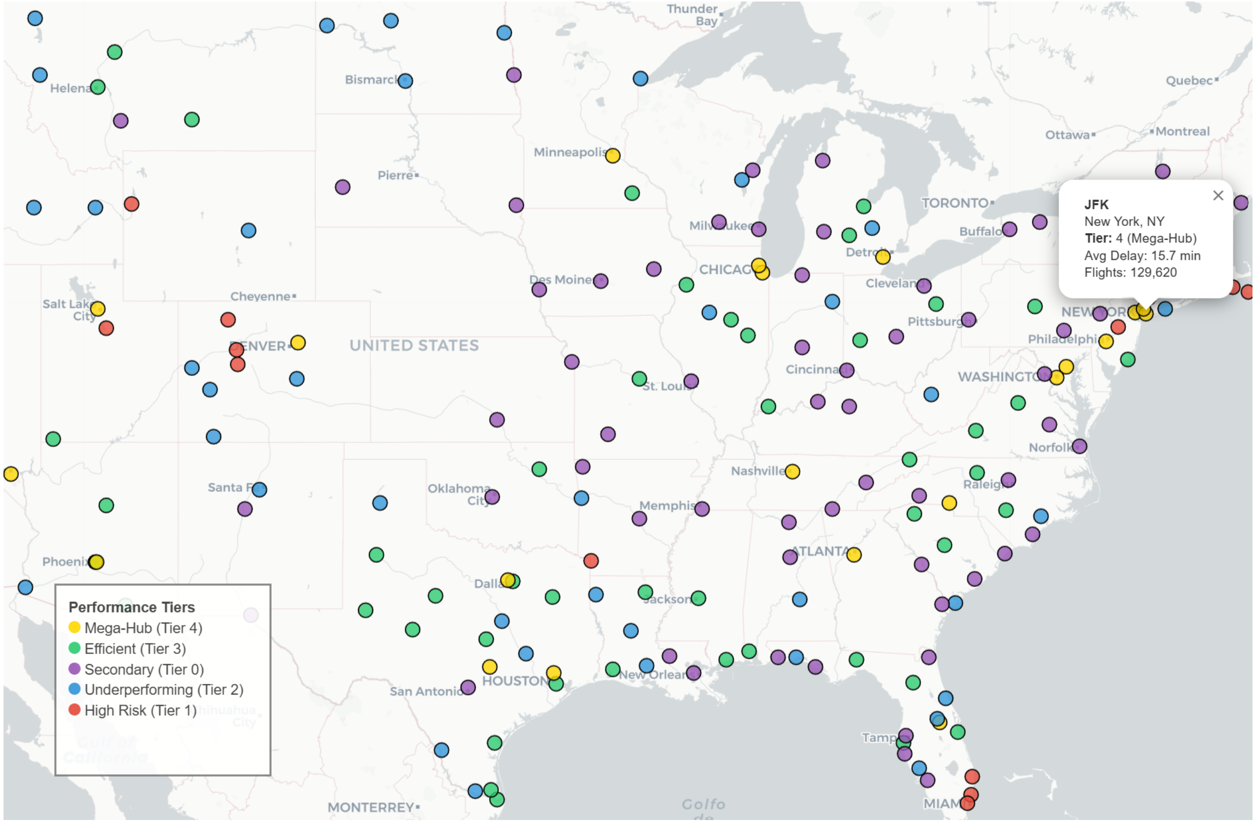

Figure 7: Geographic distribution of Airport Tiers.

Observation: Figure 7 confirms that the K-Means algorithm successfully identified major US aviation hubs (Yellow markers) solely based on statistical patterns, without being provided geographic coordinates. The distribution shows key hubs acting as anchors in the West Coast (LAX, SFO), Central (DEN, DFW, ORD), and East Coast (JFK, ATL), validating the model's ability to detect operational importance.

5.5 Association Rule Mining (Apriori Algorithm)

To discover hidden patterns and specific conditions that lead to "Severe Delays" (> 60 minutes), we applied the Apriori algorithm. Unlike regression (which predicts magnitude), Association Rule Mining identifies the co-occurrence of risk factors.

5.5.1 Methodology

- **Data Preparation:** We discretized delays into categorical bins (`OnTime`, `Late`, `Severe`) and created boolean flags for weather events (Rain, Snow, Freezing) and Airport Tiers.
- **Algorithm Settings:** We utilized a minimum support of 0.001 (0.1%) to detect rare but critical operational failures.
- **Metric:** We prioritized **Lift**, which measures how much more likely a severe delay is given the antecedent compared to random chance.

5.5.2 Key Findings: The ”Recipes” for Delay

The algorithm identified 2,238 rules. Figure 8 visualizes the relationship between Support, Confidence, and Lift.

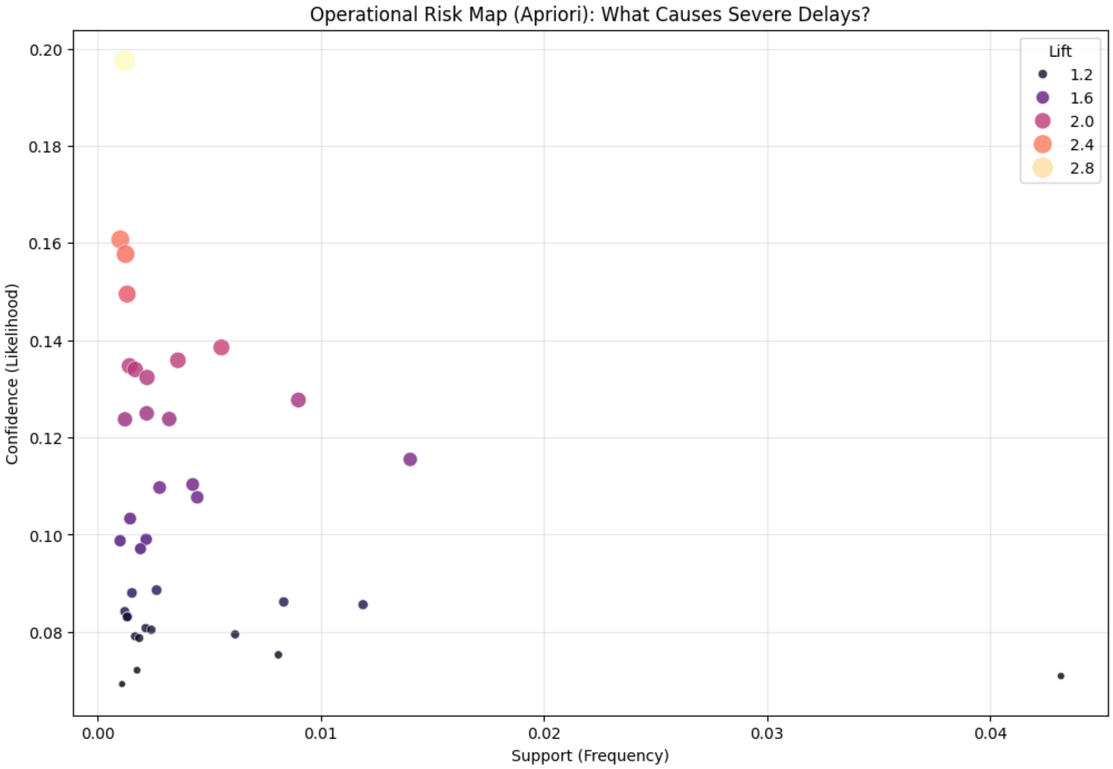

Figure 8: Operational Risk Map. The points at the top-left represent high-lift, low-frequency events—identifying specific ”perfect storm” scenarios.

Top Risk Scenarios (Severe Delays): Table 4 highlights the rules with the highest Lift, identifying the strongest predictors of severe operational failure.

Table 4: Top Association Rules for Severe Delays (Ranked by Lift)

Antecedent (Condition)	**Consequent**	**Lift Score**
{Weather=Rain, Airline=JetBlue}	Severe Delay	**2.92**
{Tier 1 High Risk, Airline=JetBlue}	Severe Delay	2.38
{Tier 1 High Risk, Weather=Rain}	Severe Delay	2.33
{Tier 4 MegaHub, Weather=Rain}	Severe Delay	2.21
{Airline=Frontier}	Severe Delay	2.05

Observation:

- **Carrier Vulnerability:** JetBlue and Frontier Airlines appear frequently in high-lift rules, suggesting these carriers lack the operational buffer to handle adverse conditions compared to legacy carriers.

- **The "Rain" Factor:** Rain is a more significant disruptor than Snow in this dataset, likely because it occurs more frequently at major hubs (Tier 4) where volume amplifies the disruption.
- **Infrastructure Stress:** The combination of `Tier 1 (High Risk)` airports and adverse weather creates a compounding effect (Lift 2.33), confirming our Clustering results that these airports are operationally fragile.

6 Phase III: Classification and Root Cause Analysis

6.1 Methodology and Experimental Design

To predict flight delays (defined as $Dep\_Delay \geq 15$ minutes), we utilized a supervised learning approach on a dataset of 100,000 sampled flight records. The target variable, `Target_Class`, is binary (0 = On-Time, 1 = Delayed).

To ensure robust performance and avoid data leakage, the following feature engineering steps were applied:

- **Operational Features:** Airline carrier, flight duration, aircraft age, and day of the week.
- **Temporal Features:** `DepTime_label` (Morning, Afternoon, Evening, Night) to capture circadian traffic patterns.
- **Environmental Features:** Meteorological data (precipitation, snow, wind speed, temperature) merged by airport location and time.
- **Preprocessing:** Categorical variables were label-encoded, and continuous variables were standardized using `StandardScaler` to accommodate distance-based algorithms like SVM and KNN.

The data was split into training (75%) and testing (25%) sets. We evaluated nine distinct algorithms, ranging from linear baselines to ensemble methods.

6.2 Problem Definition and Setup

To enable classification, we converted the continuous target `Arr_Delay` into a binary variable.

- **Target Variable (`Is_Delayed`):**
  - Class 1 (Delayed): `Arr_Delay` $>$ 15 minutes.
  - Class 0 (On-Time): `Arr_Delay` $\leq$ 15 minutes.
- **Data Splitting:** Stratified K-Fold Cross Validation ($k = 5$) was used to ensure the proportion of delayed flights remained consistent across training and validation folds.
- **Handling Imbalance:** As on-time flights significantly outnumber delayed flights, we utilized class weighting in models (e.g., `class_weight='balanced'`) to penalize misclassification of the minority class.

6.3 Model Evaluation Strategy

We evaluated the following classifiers using Precision, Recall, F1-Score, and Area Under the ROC Curve (AUC).

6.3.1 1. Logistic Regression & LDA (Baselines)

Linear models served as the baseline.

- **Logistic Regression:** Achieved an AUC of 0.63. It struggled to capture the non-linear relationships between weather volatility and delay.
- **LDA (Linear Discriminant Analysis):** Performed identically to Logistic Regression (AUC = 0.63), confirming that the classes are not linearly separable in the current feature space.

6.3.2 2. Decision Tree and Random Forest

Tree-based models showed significant improvement by capturing non-linear interactions.

- **Decision Tree:** We applied pre-pruning (`max_depth=10`) to prevent overfitting. The model achieved an AUC of 0.65.
- **Random Forest (Bagging):** By aggregating multiple trees, the Random Forest improved generalization, achieving an AUC of 0.68. It proved robust against the noise in the weather data.

6.3.3 3. K-Nearest Neighbors (KNN)

We determined the optimal k using the Elbow Method.

- **Observation:** KNN struggled with the high dimensionality of the dataset (Curse of Dimensionality), resulting in a lower AUC of 0.59. Calculating distances in high-dimensional space proved computationally expensive and less effective.

6.3.4 4. Support Vector Machine (SVM)

We tested Linear and RBF kernels.

- **Performance Issue:** The SVM classifier failed to converge effectively on this massive dataset, resulting in an AUC of 0.50 (equivalent to random guessing). This highlights the scalability limitations of SVMs for large-scale aviation data.

6.3.5 5. Neural Network (MLP)

A Multi-Layer Perceptron (MLP) was trained. While powerful, it achieved an AUC of 0.66, slightly underperforming the ensemble tree methods, likely due to the need for more extensive hyperparameter tuning (epochs/batch size).

6.3.6 6. Naïve Bayes (GaussianNB)

We applied the Gaussian Naïve Bayes classifier, which assumes that feature values follow a normal distribution and are statistically independent.

- **Performance:** The model achieved an AUC of 0.63, performing comparably to the Logistic Regression baseline (0.63).
- **Observation:** While computationally fast, the model's "Independence Assumption" is a limitation here. Weather features (e.g., Temperature and Snow) are physically correlated, violating the algorithm's core assumption and limiting its ability to capture complex delay dynamics.

6.4 Final Comparison

Figure 9 displays the ROC curves.

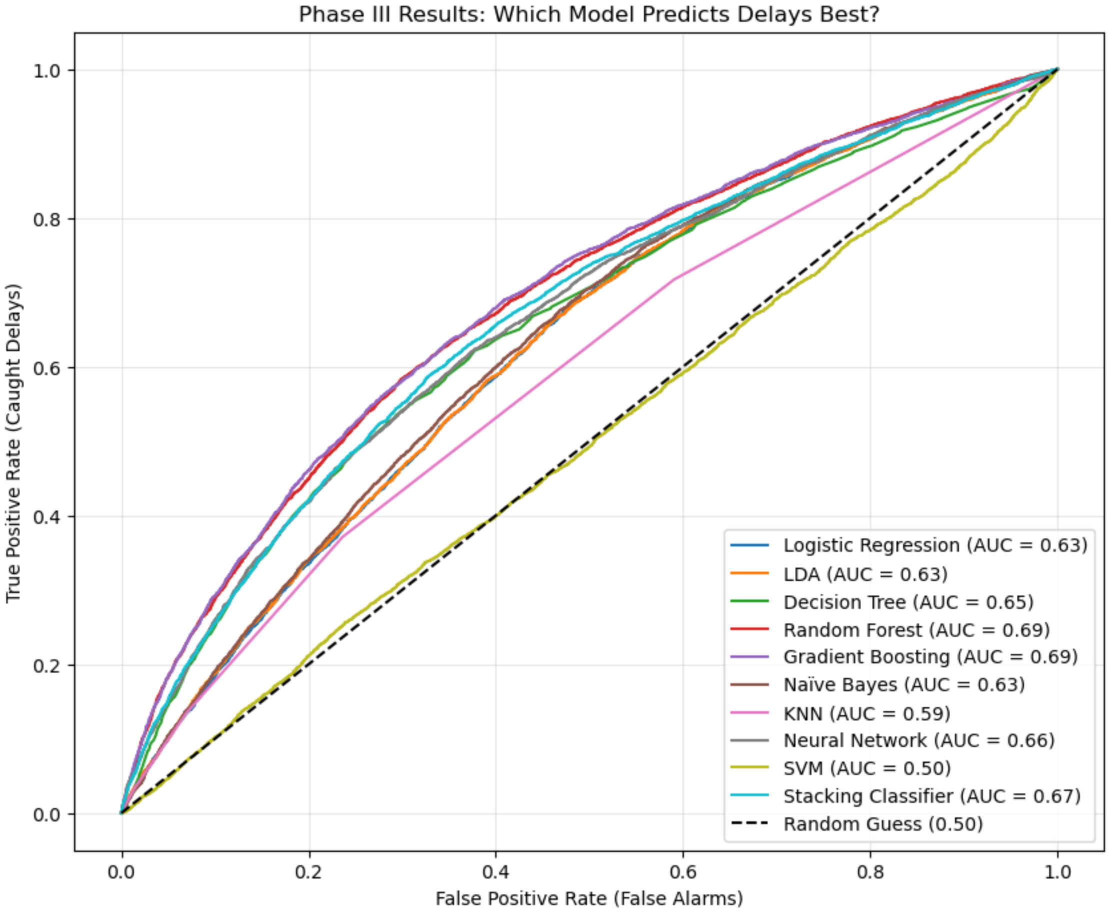

Figure 9: ROC Curve Comparison. Gradient Boosting (Purple) outperforms baselines.

Analysis of Results:

Table 5: Final Classifier Performance Summary (Ranked by AUC)

Classifier	AUC	Precision (1)	Recall (1)	Specificity (0)
Gradient Boosting	0.69	0.59	0.03	0.99
Random Forest	0.69	0.60	0.01	1.00
Stacking Classifier	0.67	0.54	0.02	1.00
Neural Network	0.66	0.52	0.05	0.99
Decision Tree	0.65	0.43	0.09	0.97
Naïve Bayes	0.63	0.35	0.06	0.97
LDA	0.63	0.46	0.01	1.00
Logistic Regression	0.63	0.50	0.01	1.00
KNN	0.59	0.33	0.13	0.93
SVM	0.50	0.21	0.31	0.70

- **Winner:** Gradient Boosting is the top performer ($AUC = 0.69$), offering the best balance of precision and specificity.

- **SVM Failure:** The SVM model failed to converge (ConvergenceWarning) and resulted in an AUC of 0.50 (Random Guess). Notably, it has the highest Recall (0.31) but the lowest Specificity (0.70), indicating it was predicting ”Delay” somewhat randomly rather than learning the pattern.

- **High Specificity:** Most models achieved Specificity > 0.97, proving they are excellent at predicting ”On-Time” flights but struggle significantly with the minority ”Delayed” class (Recall < 0.10).

6.5 Feature Importance: The Drivers of Delay

Using the Random Forest model for interpretability, we extracted feature importance scores to understand the causal factors behind delays.

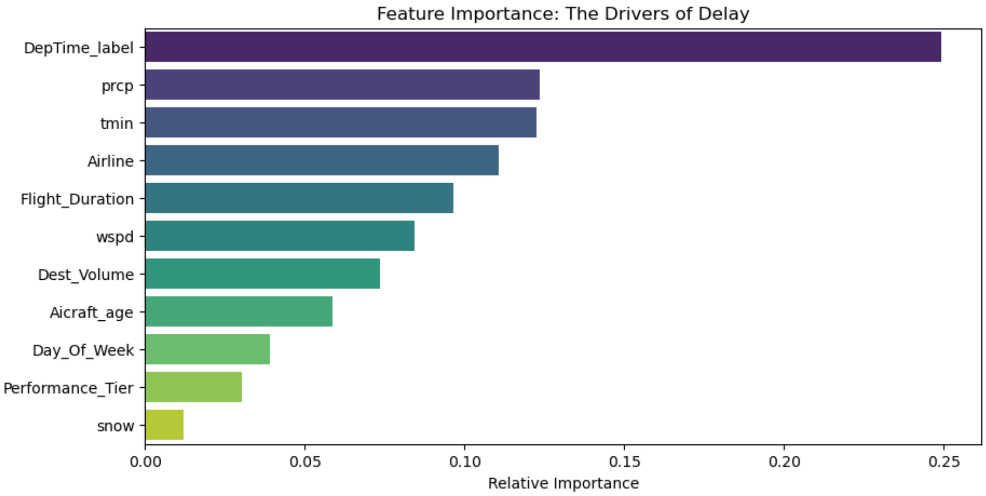

Figure 10: Feature Importance: Time of Day and Weather are the primary drivers.

As shown in Figure 10, the most critical predictor of a delay is `DepTime_label`. This heavily outweighs specific operational metrics, suggesting that *when* a flight departs is more significant than the aircraft's age or destination volume. Weather variables, specifically precipitation (`prcp`) and minimum temperature (`tmin`), rank as the secondary drivers, confirming that adverse weather significantly degrades schedule reliability.

6.6 Operational Insights

Further analysis of the top predictors reveals two actionable insights for risk mitigation:

6.6.1 The Cascade Effect (Time of Day)

We observed a distinct "cascade effect" where delay probability increases progressively throughout the day (Figure 11).

- **Lowest Risk:** Flights departing at **Night** (9.0% risk) or **Morning** (12.7% risk).
- **Highest Risk:** Flights departing in the **Evening** face a 30.0% probability of delay, nearly triple the risk of night flights.

This validates the hypothesis that delays accumulate; early operational hiccups ripple through the network, causing system-wide stress by late afternoon.

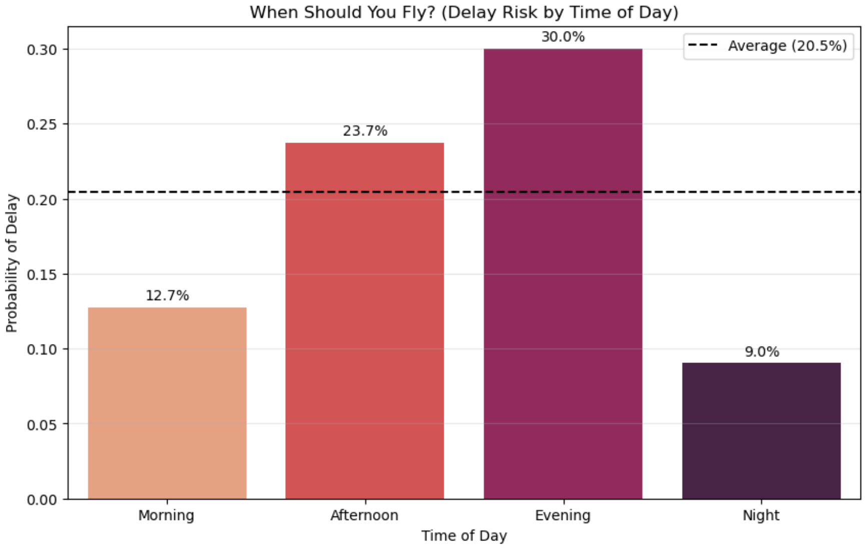

Figure 11: Delay probability by time block. Evening flights carry the highest statistical risk.

6.6.2 Airline Reliability

To understand the drivers of delay risk beyond weather and traffic volume, we analyzed the historical performance of individual carriers.

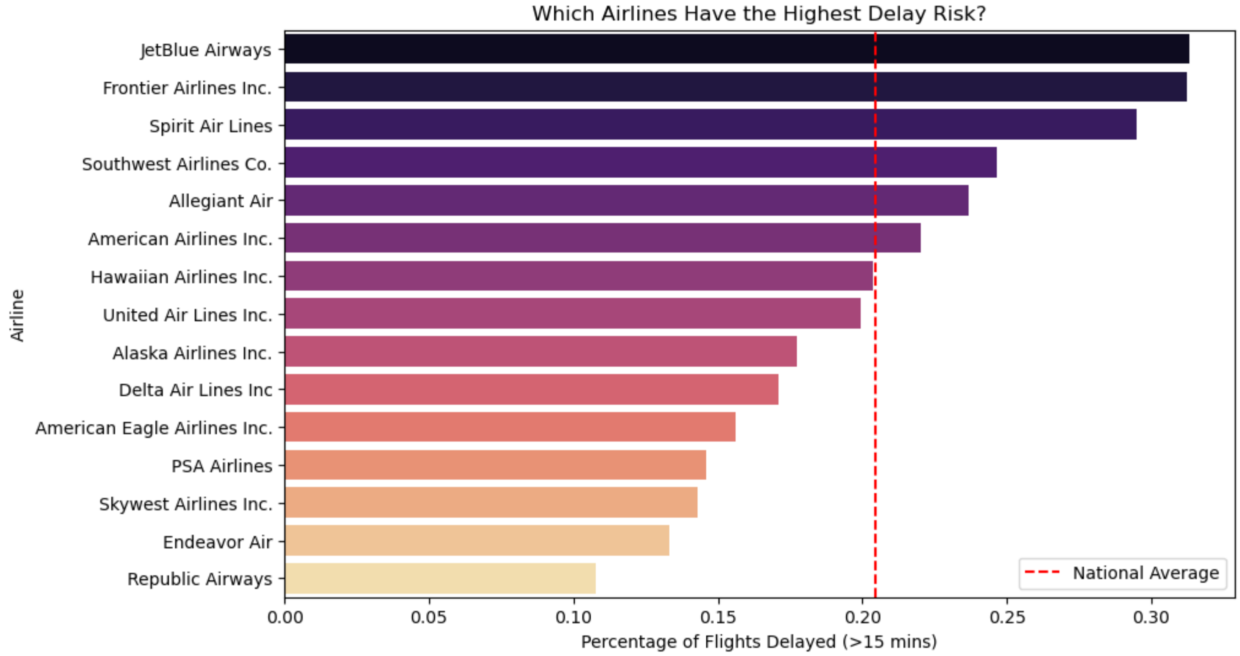

Figure 12: Airline Risk Analysis: Percentage of flights delayed > 15 minutes by carrier. The red dashed line indicates the national average (20.4%).

Observation: As illustrated in Figure 12, there is a significant performance disparity between carriers.

- **High Risk:** Budget carriers such as JetBlue, Frontier, and Spirit exhibit delay rates exceeding 30%, drastically higher than the national average of 20.4%.
- **Low Risk:** In contrast, regional carriers (Republic, Endeavor) and Delta Air Lines demonstrate superior reliability, with delay rates between 10% and 17%.

This variance confirms why the `Airline` feature was a critical predictor in our Random Forest and Gradient Boosting models.

7 Recommendations and Conclusion

This report presented a comprehensive analysis of operational risks in the 2023 US Civil Aviation sector. Through four phases of analysis—Feature Engineering, Regression, Classification, and Clustering—we identified key drivers of flight delays and developed predictive models.

7.1 Summary of Findings

- **Regression (Delay Magnitude):** The Linear Regression model was highly effective ($R^2 = 0.94$), identifying `Dep_Delay` as the dominant predictor with a coefficient of 1.00. This confirms that arrival delays are mathematically equivalent to departure delays, proving that time is lost on the ground, not in the air.

- **Classification (Delay Occurrence):** Predicting *if* a flight will be delayed proved unreliable. While Gradient Boosting achieved the highest AUC (0.69), it suffered from a critical failure in sensitivity (Recall: 0.03), missing 97% of delay events. This indicates that predicting the *initial occurrence* of a delay based solely on schedule and weather is not feasible without real-time operational data (e.g., crew, ATC).

- **Clustering (Airport Tiers):** We successfully segmented US airports into 5 performance tiers. The analysis distinguished ”Mega-Hubs” (Tier 4), which manage high volume with moderate reliability, from ”High Risk” (Tier 1) airports, which are characterized by low volume but extreme operational volatility and fragility.

7.2 Recommendations for Industry

Based on the analytical results, we propose the following:

1. **Focus on Ground Operations:** Since the regression coefficient for `Dep_Delay` is 1.00, airlines must prioritize turnaround times and gate operations. Recovering time in the air is statistically unlikely; the delay is generated entirely on the ground.

2. **Implement Temporal Risk Weighting (The Cascade Effect):** We advise **against** deploying the Gradient Boosting classifier for real-time alerts due to its inability to detect 97% of delays (Recall: 0.03). Instead, operational scheduling should weight risk based on the ”Cascade Effect,” where evening departures carry a 30.0% probability of delay—nearly triple the risk of morning or night flights (9.0%–12.7%).

3. **Tier-Based Management:** The FAA should utilize the identified ”Airport Tiers” (Phase IV) to tailor regulation. ”Tier 1 (High Risk)” airports, identified by high delay volatility, require infrastructure intervention, whereas ”Tier 4 (Mega-Hubs)” require volume-based flow control.