SkyStream

Air Traffic Delay Prediction
Integrating Economic Cost Analysis

Research Problem

Importance of Predicting Air Traffic Delays

Economic Impact of Flight Delays

Flight delays impose substantial economic costs, including lost revenue for airlines, increased operational expenses, and customer compensation claims. The overall impact can reach billions annually, stressing the need for precise predictions.

Limitations of Traditional Classification Methods

Traditional classification methods often rely on binary outcomes, failing to account for the complex patterns of flight delays. These methods lack the precision needed for effective operational adjustments, leading to inadequate responses to delay scenarios.

Importance of Predicting Air Traffic Delays

- Flight delays are more than just an inconvenience — they ripple across the entire economy.
- In 2022, flight delays and cancellations affected 200 million passengers in the US.
- Over 650 million hours of passenger time were lost due to disruptions (DOT, 2022).
- The total economic loss due to delays was estimated at $30–34 billion.
- Our project investigates:
  - What causes delays?
  - Can we predict them early?
  - What is their financial impact?

Limitations of Traditional Classification Methods

Traditional classification methods often rely on binary outcomes, failing to account for the complex patterns of flight delays. These methods lack the precision needed for effective operational adjustments, leading to inadequate responses to delay scenarios.

Economic Impact of Flight Delays

Flight delays impose substantial economic costs, including lost revenue for airlines, increased operational expenses, and customer compensation claims. The overall impact can reach billions annually, stressing the need for precise predictions.

Dataset and Data Exploration

Overview of Historical U.S. Flight Performance Data

The Kaggle dataset includes records of delays, cancellations, and attributing factors for prediction modeling.

- compiled from Bureau of Transportation
- 3+ GB of flight performance data
- 10M+ distinct flights over the past few years
- 20+ different US airlines

Key Variables and Features Explored

These features are critical as they provide insights into patterns and trends that affect flight delays, allowing for more accurate predictions.

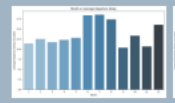
- FlightDate
- Airline
- Origin, Dest
- DepDelayMinutes, ArrDelayMinutes, Distance, AirTime

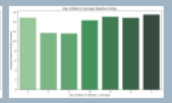
Generated synthetic data to approximate flight size based on distance & airline revenue loss per minute of delay, based on FAA and Airlines for America figures ($74-$100/min).

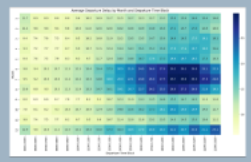
- EstimatedPassengers
- Plane Type
- AvgTicketPrice
- EstimatedLossUSD

- Distributed Data Processing with PySpark:
  - Spark was used to load, clean, transform, and filter millions of flight records.
  - Handled null values, outlier removal, feature engineering, and label creation at scale
- Why Spark Was Essential:
  - Enabled parallel processing on large files without crashing.
  - Allowed us to scale feature engineering and model data compilation quickly and efficiently.

Time-Series Analysis







Most delays occur:
- Summer months
- End of the week
- Between 6 - 11pm

due to peak demand, cascading delays, constrained resources, weather-related vulnerabilities

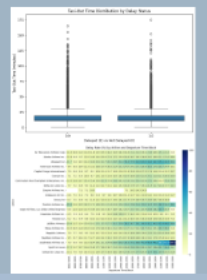
Insights from EDA

Taxi-Out Time Distribution by Delay Status

- A boxplot comparing taxi-out times (the time from gate pushback to takeoff) for flights that were delayed vs not delayed
- Delayed flights tend to have slightly longer taxi-out times on average.
- However, both delayed and non-delayed flights show significant overlap in distribution.
- Presence of outliers in both groups hints at sporadic operational bottlenecks (e.g., airport congestion).

Airline vs Departure Time Block – Delay Rate (%)

- A heatmap of average delay rate by airline and departure time block, revealing when each airline tends to experience higher delays.
- Airlines like JetBlue and Southwest experience higher delays during evening blocks (6:00 PM – 11:59 PM).
- Some carriers (e.g., Alaska, Horizon) maintain low delay rates across most time blocks.
- This suggests that both airline operations and schedule timing contribute to delay patterns.



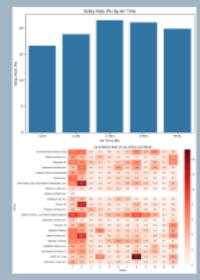
More Insights

Delay Rate (%) by Air Time

- Flights in the 2–3 hour range experience the highest delay rate (~22%).
- Flights over 3 hours also show high delays, but slightly lower than 2–3 hour ones.
- Flights under 1 hour are least prone to delays (~16%).

Cancellation Rate (%) by Airline and Month

- Airlines like Spirit, Mesa, and GoJet show significantly higher cancellation rates, especially in Jan, Feb, and Sep.
- Cancellations spike in early and late months, possibly due to seasonal disruptions.
- Some airlines like Delta and Hawaiian maintain consistently low cancellation rates year-round.



Overview of Historical U.S. Flight Performance Data

The Kaggle dataset includes records of delays, cancellations, and attributing factors for prediction modeling.

- compiled from Bureau of Transportation
- 3+ GB of flight performance data
- 10M+ distinct flights over the past few years
- 20+ different US airlines

Key Variables and Features Explored

These features are critical as they provide insights into patterns and trends that affect flight delays, allowing for more accurate predictions.

- FlightDate
- Airline
- Origin, Dest
- DepDelayMinutes, ArrDelayMinutes, Distance, AirTime

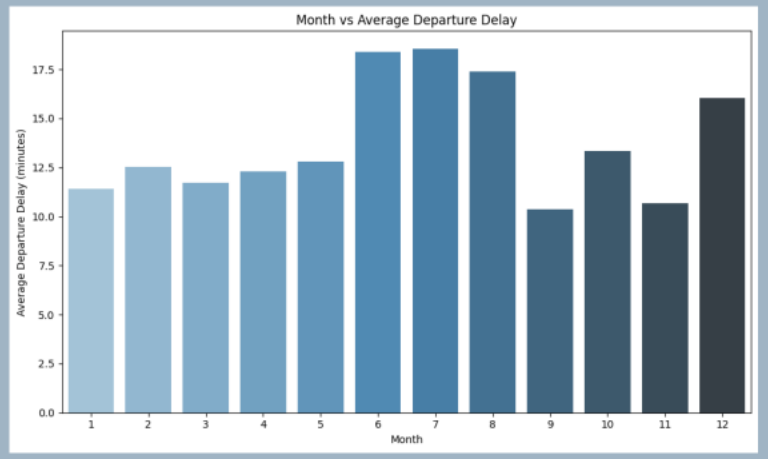
Generated synthetic data to approximate flight size based on distance & airline revenue loss per minute of delay, based on FAA and Airlines for America figures ($74–$100/min).

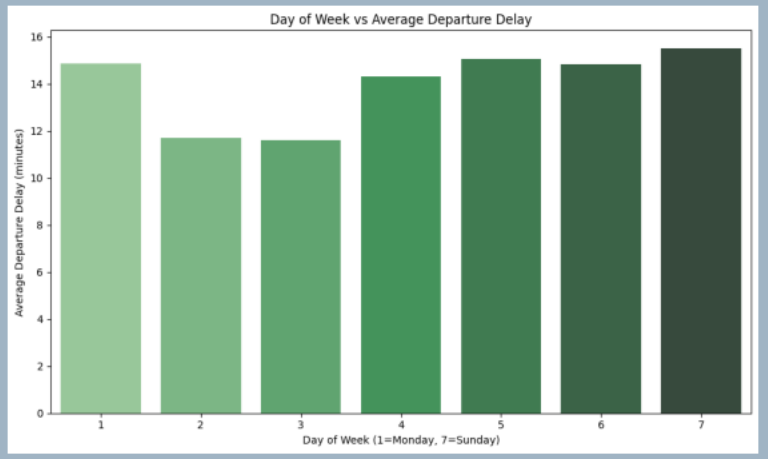
- EstimatedPassengers
- Plane Type
- AvgTicketPrice
- EstimatedLossUSD

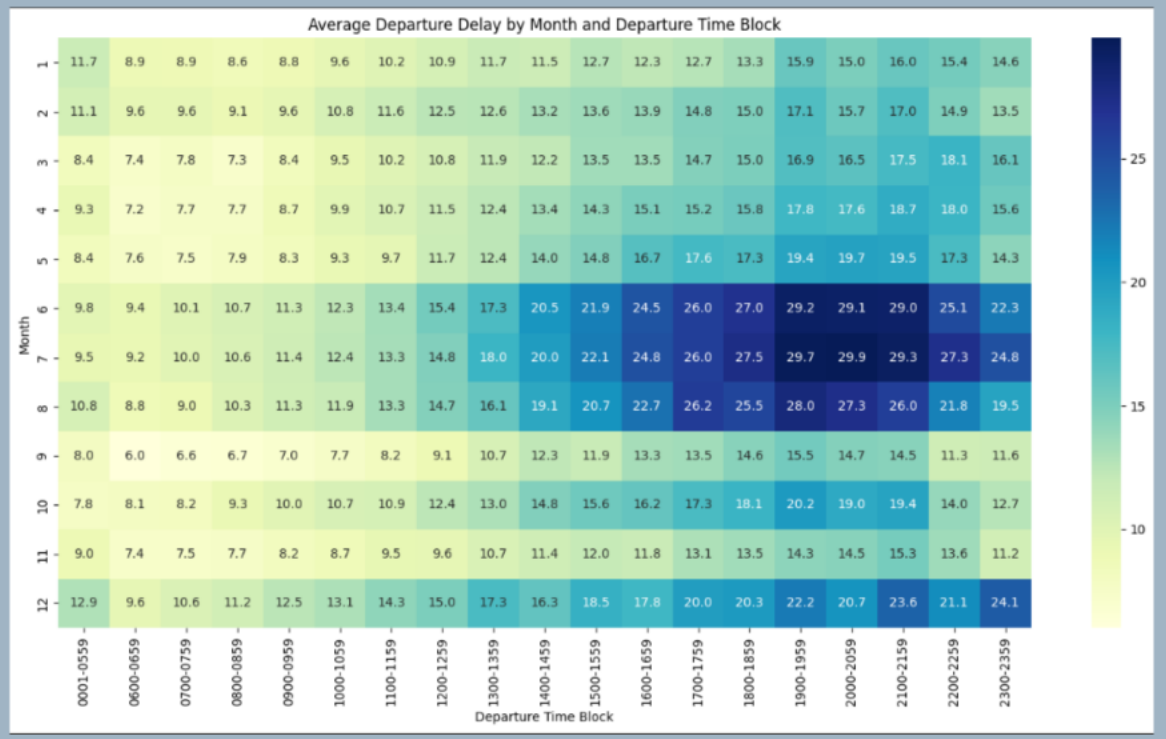
- Distributed Data Processing with PySpark:
  - Spark was used to load, clean, transform, and filter millions of flight records.
  - Handled null values, outlier removal, feature engineering, and label creation at scale

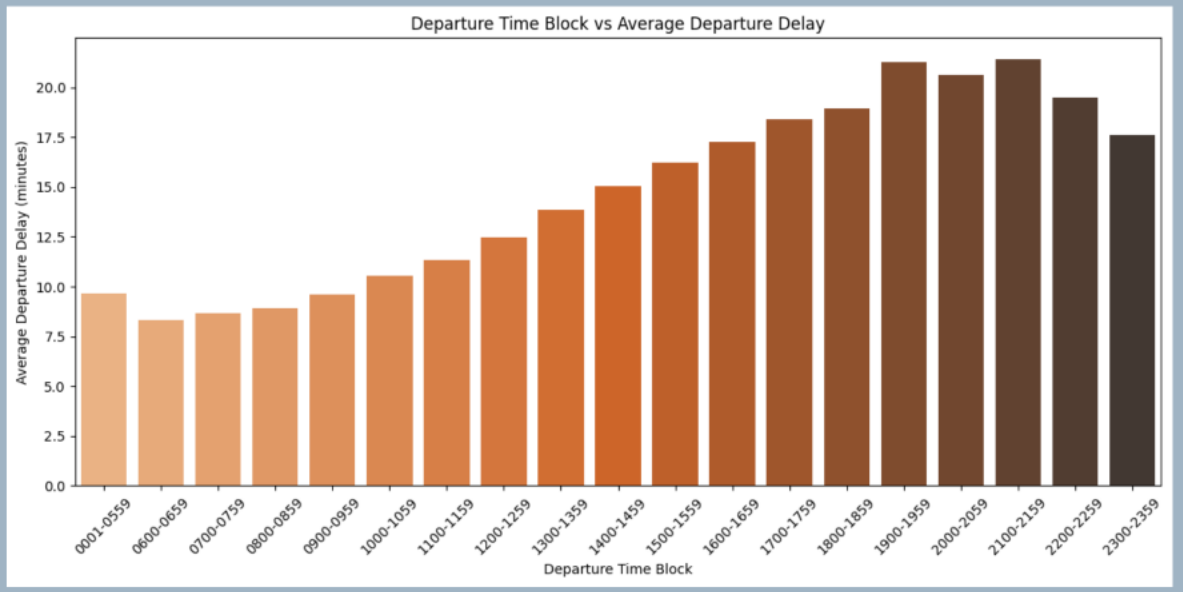
- Why Spark Was Essential:
  - Enabled parallel processing on large files without crashing.
  - Allowed us to scale feature engineering and model data compilation quickly and efficiently.

Time-Series Analysis









Most delays occur:

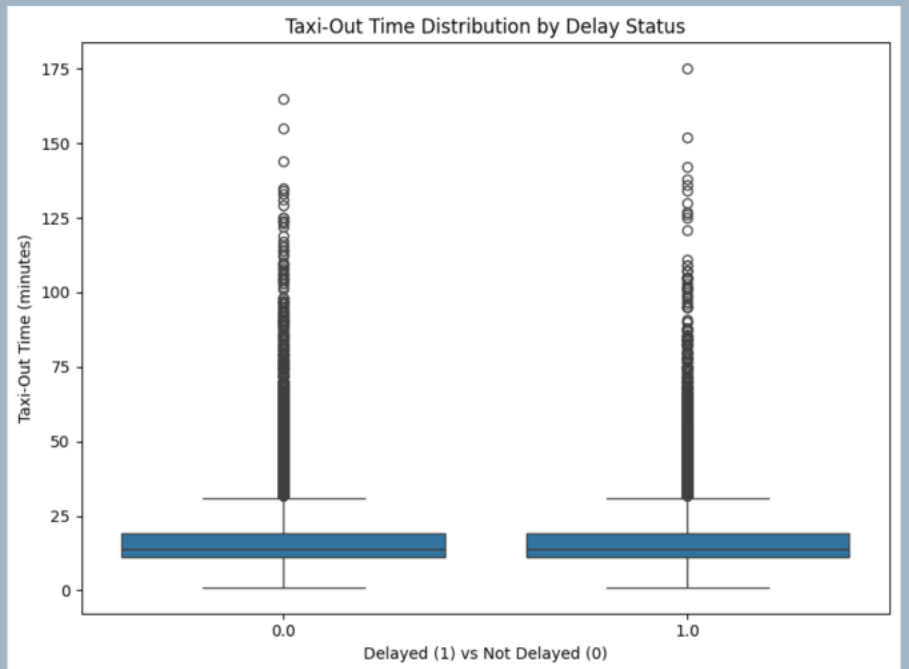
- Summer months
- End of the week
- Between 6 - 11pm

due to peak demand, cascading delays, constrained resources, weather-related vulnerabilities

Insights from EDA

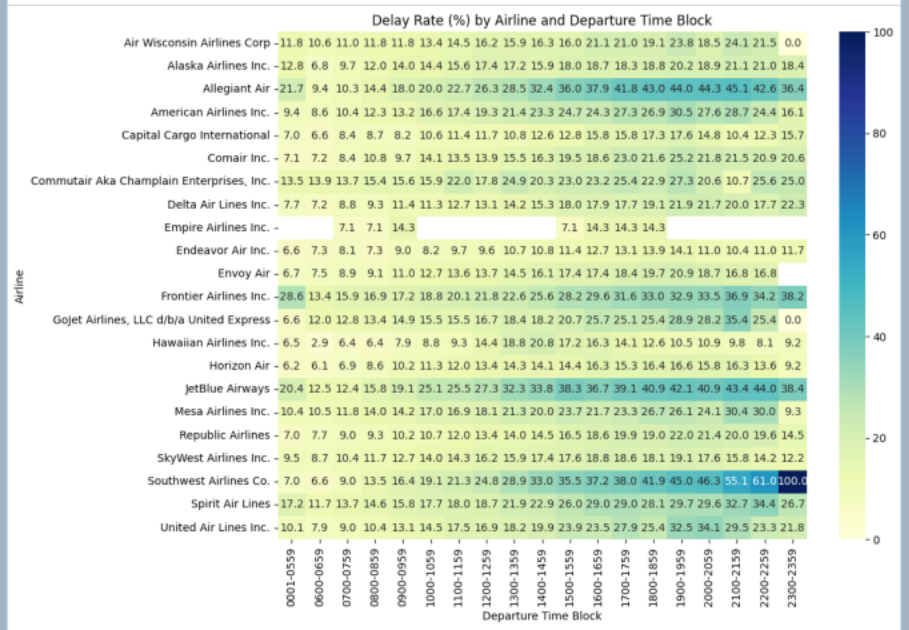
Taxi-Out Time Distribution by Delay Status

- A boxplot comparing taxi-out times (the time from gate pushback to takeoff) for flights that were delayed vs not delayed
- Delayed flights tend to have slightly longer taxi-out times on average.
- However, both delayed and non-delayed flights show significant overlap in distribution.
- Presence of outliers in both groups hints at sporadic operational bottlenecks (e.g., airport congestion).



Airline vs Departure Time Block – Delay Rate (%)

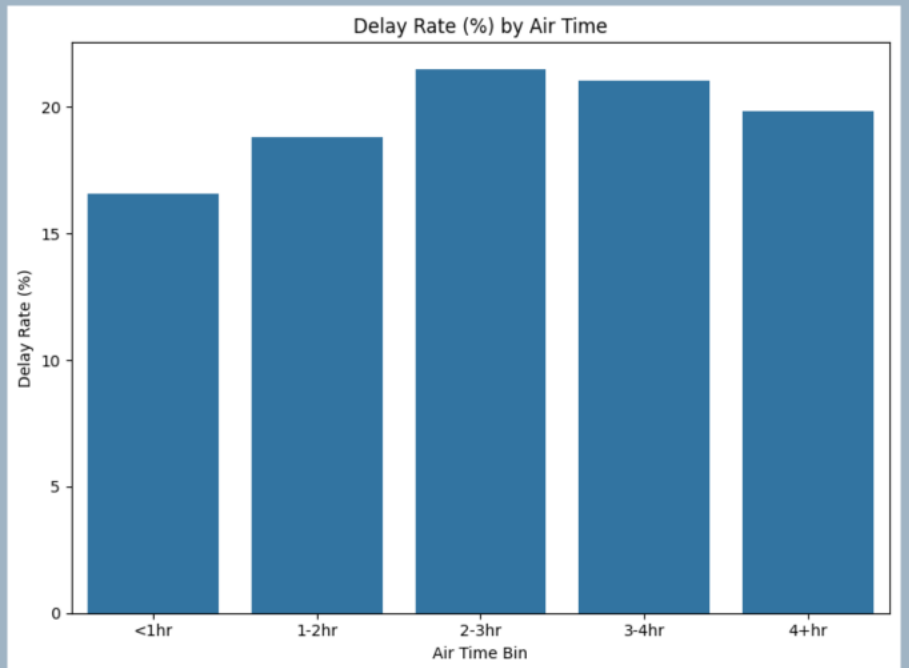
- A heatmap of average delay rate by airline and departure time block, revealing when each airline tends to experience higher delays.
- Airlines like JetBlue and Southwest experience higher delays during evening blocks (6:00 PM – 11:59 PM).
- Some carriers (e.g., Alaska, Horizon) maintain low delay rates across most time blocks
- This suggests that both airline operations and schedule timing contribute to delay patterns.



More Insights

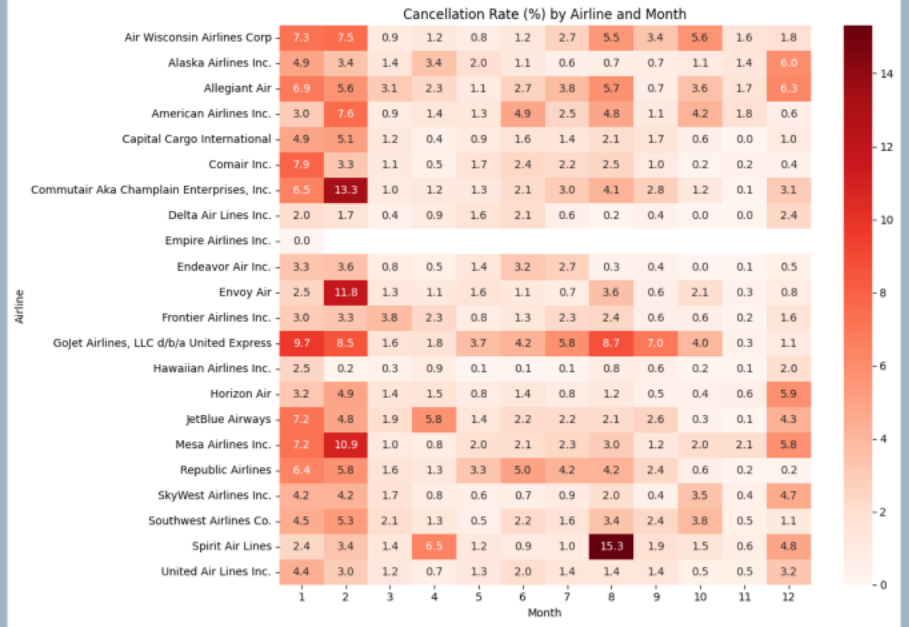
Delay Rate (%) by Air Time

- Flights in the 2–3 hour range experience the highest delay rate (~22%).
- Flights over 3 hours also show high delays, but slightly lower than 2–3 hour ones.
- Flights under 1 hour are least prone to delays (~16%).



Cancellation Rate (%) by Airline and Month

- Airlines like Spirit, Mesa, and GoJet show significantly higher cancellation rates, especially in Jan, Feb, and Sep.
- Cancellations spike in early and late months, possibly due to seasonal disruptions.
- Some airlines like Delta and Hawaiian maintain consistently low cancellation rates year-round.



Technical Solutions and Results

Economic Imapct on Airlines

- We employed the DecisionTreeRegressor from PySpark's pyspark.ml.regression module to analyze airline delays' financial impact.
- The model forecasts economic loss in USD, utilizing average flight delays as key predictors.
- Our comparison of actual vs. predicted losses included major regional airlines such as Southwest Airlines Co. and GoJet Airlines.
- An R-squared (R^2) value of 0.8270 demonstrates a strong correlation between predicted and actual losses.
- This high R^2 value confirms the model's effectiveness in capturing data variance.

```
==================================================
+--------------------------------------+-----------------+------------+---------------+----------------+
|Airline                               |AvgDelay         |ActualLossUSD|PredictedLossUSD|CostPerMinuteUSD|
+--------------------------------------+-----------------+------------+---------------+----------------+
|Southwest Airlines Co.                |36.83985765124555|10,485.49   |10,407.67      |282.51          |
|GoJet Airlines, LLC d/b/a United Express |63.28            |14,949.82   |16,208.11      |256.13          |
|Air Wisconsin Airlines Corp           |28.0             |6,662.23    |6,806.44       |243.09          |
|Commutair Aka Champlain Enterprises, Inc.|68.67857142857143|16,469.04   |14,252.51      |207.52          |
+--------------------------------------+-----------------+------------+---------------+----------------+


🔍 Model Accuracy Metrics
==================================================
R-squared (R²): 0.8270
```

Delay Prediction

- We evaluated three classification models: **Decision Tree, Random Forest, and Gradient-Boosted Tree**, to predict flight delays using features like route, time of day, and flight characteristics.

- The **Gradient-Boosted Tree** model achieved the highest performance metrics, including accuracy, precision, recall, and F1 score, all exceeding 90%, and a remarkable ROC AUC of 0.97.

- **Decision Tree** and **Random Forest** models also performed well, with Random Forest showcasing balanced results and an F1 score of 0.78.
- These findings indicate that tree-based models are effective for classifying flight delays, demonstrating strong predictive power.

- **Gradient-Boosted Trees**, in particular, offer reliability for real-time decision-making in the aviation sector.

🌲 Random Forest Results
✅ ROC AUC: 0.8742617101230684
✅ Accuracy: 0.7852736118774138
✅ Precision (for delay=1): 0.765233644859813
✅ Recall (for delay=1): 0.8223360449934719
✅ F1 Score: 0.7849842416207453

✅ ROC AUC: 0.8647205112517246
✅ Accuracy: 0.8465172379133581
✅ Precision (for delay=1): 0.7999650898935242
✅ Recall (for delay=1): 0.9232473811442385
✅ F1 Score: 0.8456297333041329

✅ ROC AUC: 0.9764047528621571
✅ Accuracy: 0.9064051763053619
✅ Precision (for delay=1): 0.9062970774329617
✅ Recall (for delay=1): 0.9062970774329617
✅ F1 Score: 0.9064051763053619

Conclusion and Discussion

Obstacles & Challenges

- Performance Bottlenecks:
  - Preprocessing (cleaning, groupings, imputations) was initially slow on local machines.
  - Early attempts to compute multivariable aggregations caused kernel crashes.
- Experimentation Constraints:
  - Iterating on EDA and model tuning required careful resource management

Summary of Findings

- SkyStream utilizes machine learning to predict air traffic delays based on historical flight data.
- The model incorporates temporal features, including months and days, for improved accuracy.
- Analysis demonstrates a direct economic impact of flight delays on airlines.
- Understanding these delays helps airlines manage and mitigate financial implications.
- SkyStream provides a valuable tool for optimizing operational efficiency in air travel.

Future Research Directions

- Real-Time Integration:
  - Incorporate real-time flight data streams from APIs (FAA, ADS-B) to make dynamic delay predictions and updates.
- Live Monitoring Dashboard:
  - Deploy a cloud-hosted dashboard (using AWS/GCP + Streamlit/Power BI) to visualize delay forecasts and trigger alerts for high-risk routes or time blocks.
- Weather & Regional Factors:
  - Enhance model accuracy by integrating weather features (wind, precipitation, visibility) and regional data such as airport infrastructure or state-wise congestion patterns.

Obstacles & Challenges

- Performance Bottlenecks:
  - Preprocessing (cleaning, groupings, imputations) was initially slow on local machines.
  - Early attempts to compute multivariable aggregations caused kernel crashes.
- Experimentation Constraints:
  - Iterating on EDA and model tuning required careful resource management

Summary of Findings

- SkyStream utilizes machine learning to predict air traffic delays based on historical flight data.
- The model incorporates temporal features, including months and days, for improved accuracy.
- Analysis demonstrates a direct economic impact of flight delays on airlines.
- Understanding these delays helps airlines manage and mitigate financial implications.
- SkyStream provides a valuable tool for optimizing operational efficiency in air travel.

Future Research Directions

- Real-Time Integration:
  - Incorporate real-time flight data streams from APIs (FAA, ADS-B) to make dynamic delay predictions and updates.
- Live Monitoring Dashboard:
  - Deploy a cloud-hosted dashboard (using AWS/GCP + Streamlit/Power BI) to visualize delay forecasts and trigger alerts for high-risk routes or time blocks.
- Weather & Regional Factors:
  - Enhance model accuracy by integrating weather features (wind, precipitation, visibility) and regional data such as airport infrastructure or state-wise congestion patterns.

SkyStream

Air Traffic Delay Prediction
Integrating Economic Cost Analysis