Milestone 3

Flight Delay Prediction

Team Members:

Harsh Gupta, UFID: 21828165
Muthukumaran Ulaganathan, UFID: 28225998

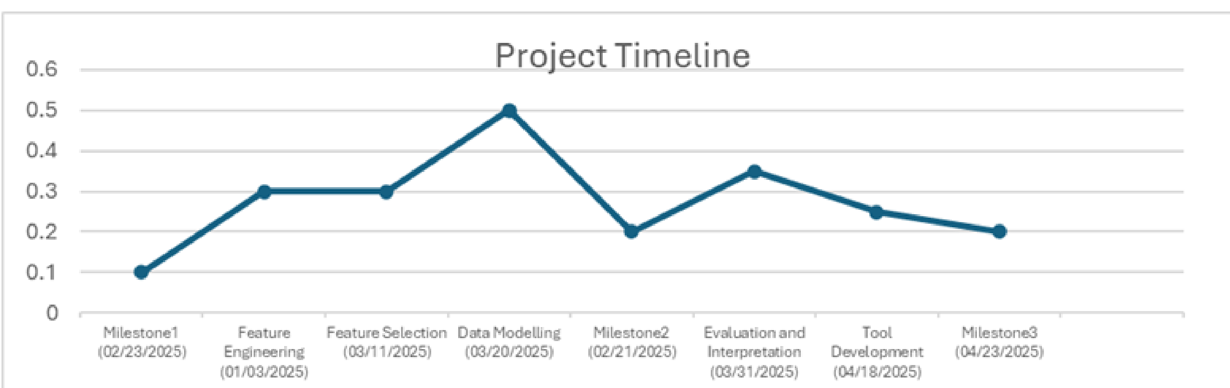
Objective:

The goal of this project is to develop an interactive dashboard that leverages machine learning models to accurately predict flight delays, using various datasets including the U.S. International Air Traffic data, US Domestic Flights, Airline codes data, Storm Events data, and Consumer Airfare Reports. By analyzing flight operations, delays, and external factors like weather events and airfare trends, the objective is to identify key contributors to flight delays and provide insights that can improve airline operational efficiency. The predictive model will be built on historical data to forecast future flight delays, enabling airlines and passengers to make more informed decisions.

Tech Stack:

- **Python**: Primary programming language
- **Pandas, NumPy, sklearn**: Libraries for data manipulation, preprocessing, and statistical analysis.
- **SQLite**: Database used..
- **Matplotlib, Seaborn, Plotly**: Tools for exploratory data analysis (EDA).
- **Scikit-Learn**: For training machine learning models.
- **Streamlit**: For building an interactive web-based dashboard and displaying model insights.

Project Timeline:

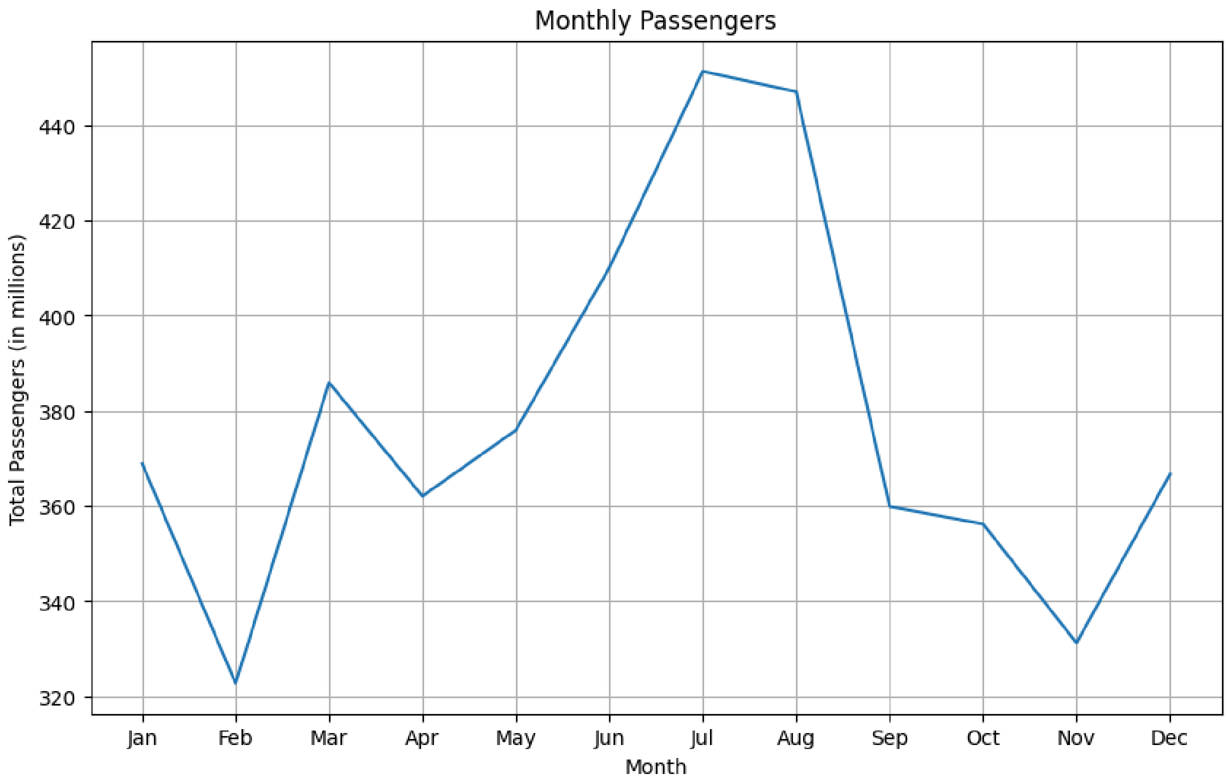


Datasets:

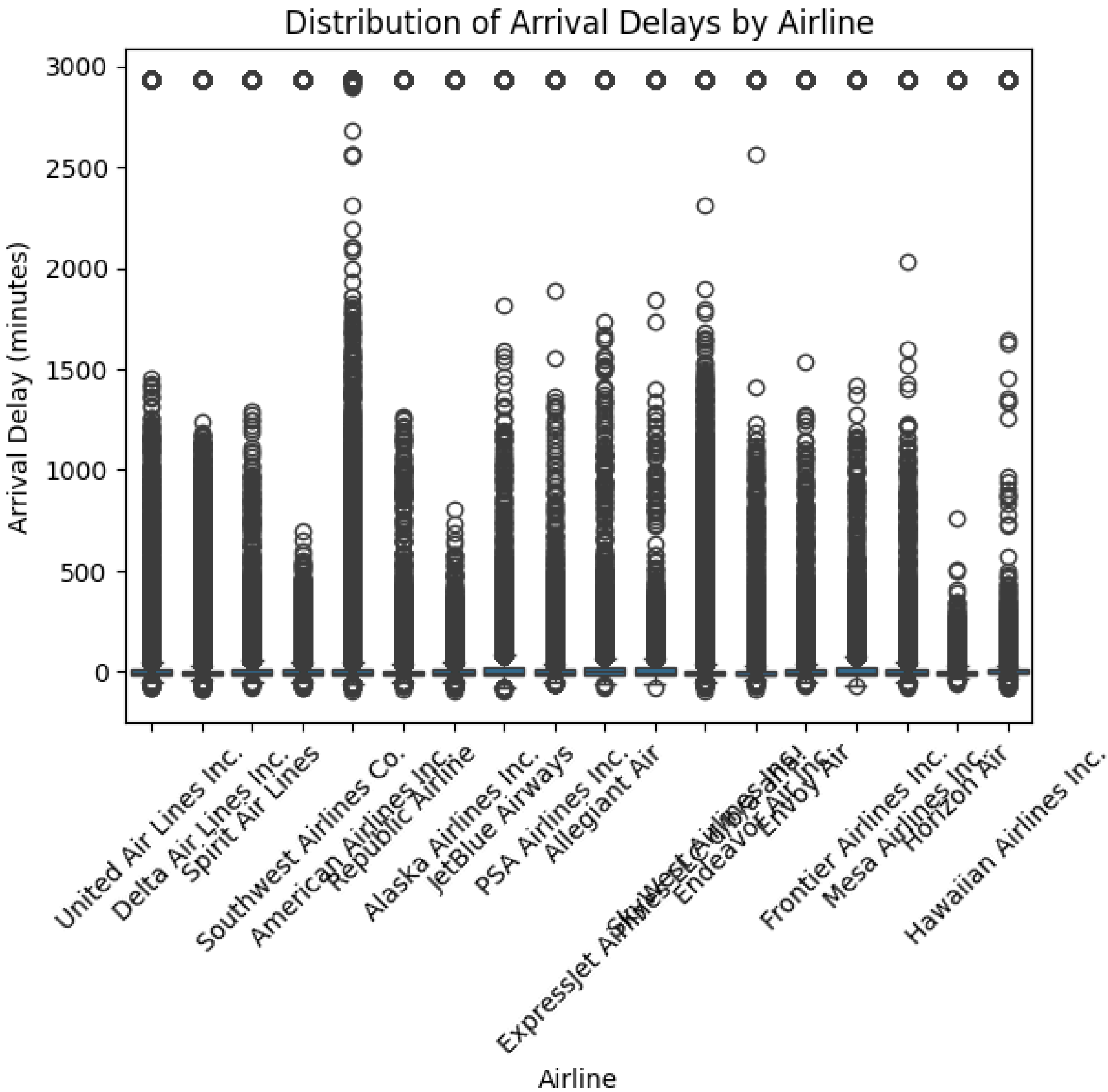
Dataset	Description	Important Columns
US International Air Traffic Data	Data on flights between US and non-US gateways.	Year, Month, US_Gateway, Foreign_Gateway, Scheduled_Flights, Charter_Flights, Passengers
Flight Delay and Cancellation (2019–2023)	Flight delay and cancellation records across 4 years.	FL_DATE, AIRLINE, FL_NUMBER, ORIGIN, DEST, DEP_DELAY, ARR_DELAY, CANCELLED
Storm Events Data	Severe weather events from NOAA covering US data since 1950.	EVENT_ID, STATE, EVENT_TYPE, BEGIN_DATE_TIME, END_DATE_TIME, DEATHS_DIRECT, DAMAGE_PROPERTY
Airline Fleets	Data on top global airlines and their fleet composition.	Parent_Airline, Airline, Aircraft_Type, Current, Future, Historic, Total, Orders
Consumer Airfare Report	Fare info for shortest (under 750 miles) high and low fare US routes.	City_Pair, Average_Fare, Passengers, Carrier, Distance, Fare_Per_Mile, Market_Rank, Quarter
List of Airline Codes	Wikipedia list of IATA, ICAO codes, and call signs for airlines worldwide.	IATA_Code, ICAO_Code, Airline_Name, Call_Sign, Country, Notes

Exploratory Data Analysis:

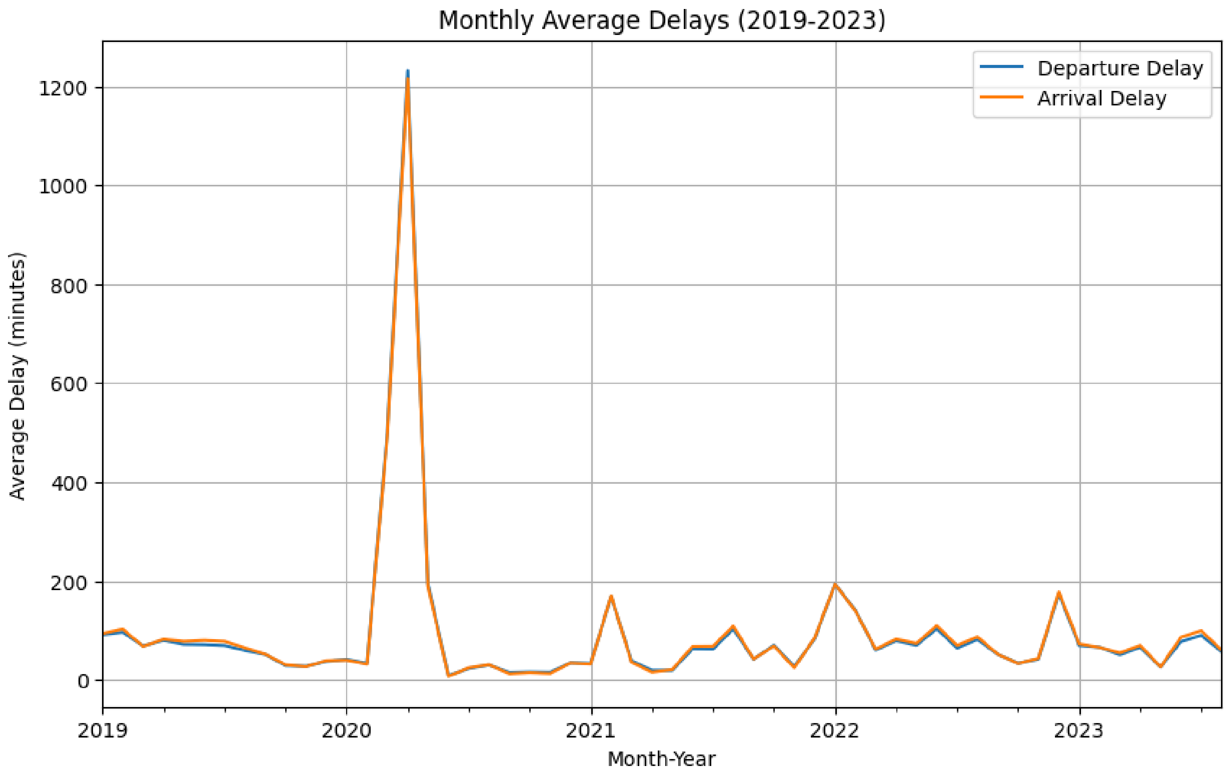
1. Passenger numbers peak during summer (July-August), indicating a seasonal travel trend. The lowest numbers are in February and November.



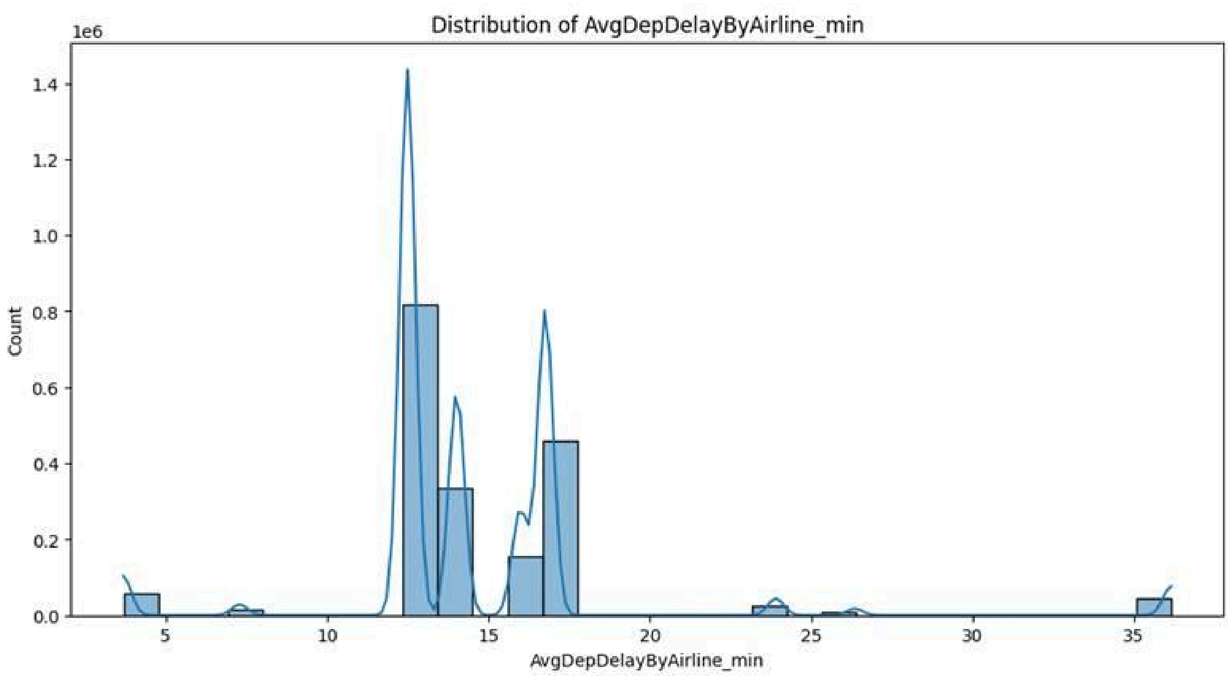
2. Most airlines experience a wide range of arrival delays, with significant outliers suggesting occasional extreme delays affecting certain carriers more than others.



3. A major spike in delays occurred in early 2020, likely due to the COVID-19 pandemic, while delays remained relatively stable in the following years with periodic fluctuations.



4. Distribution of Average Departure Delay By Airlines in minutes:



Data Integration Steps:

- Filtered the flight delay data to only include records from 2020 and pulled out the date from the timestamp.
- Merged it with international air traffic data using the IATA code and date to match delays with flight patterns.
- Added storm events data by extracting the state and joining it based on date and state to factor in weather effects.
- Cleaned the airline fleet data by dropping rows with missing key numbers, filling missing text fields with "Unknown", and replacing NaNs with 0 where needed. Also converted cost fields to numbers and made delay values positive.
- Linked the fleet data with airline codes using the airline name to get consistent IATA codes.
- Finally, combined everything with airfare data using IATA and state to build a complete dataset that includes delays, weather, fleet info, traffic, and fares.

Feature Engineering Steps:

We engineered a set of new features to improve prediction performance and reveal deeper trends in flight delays.

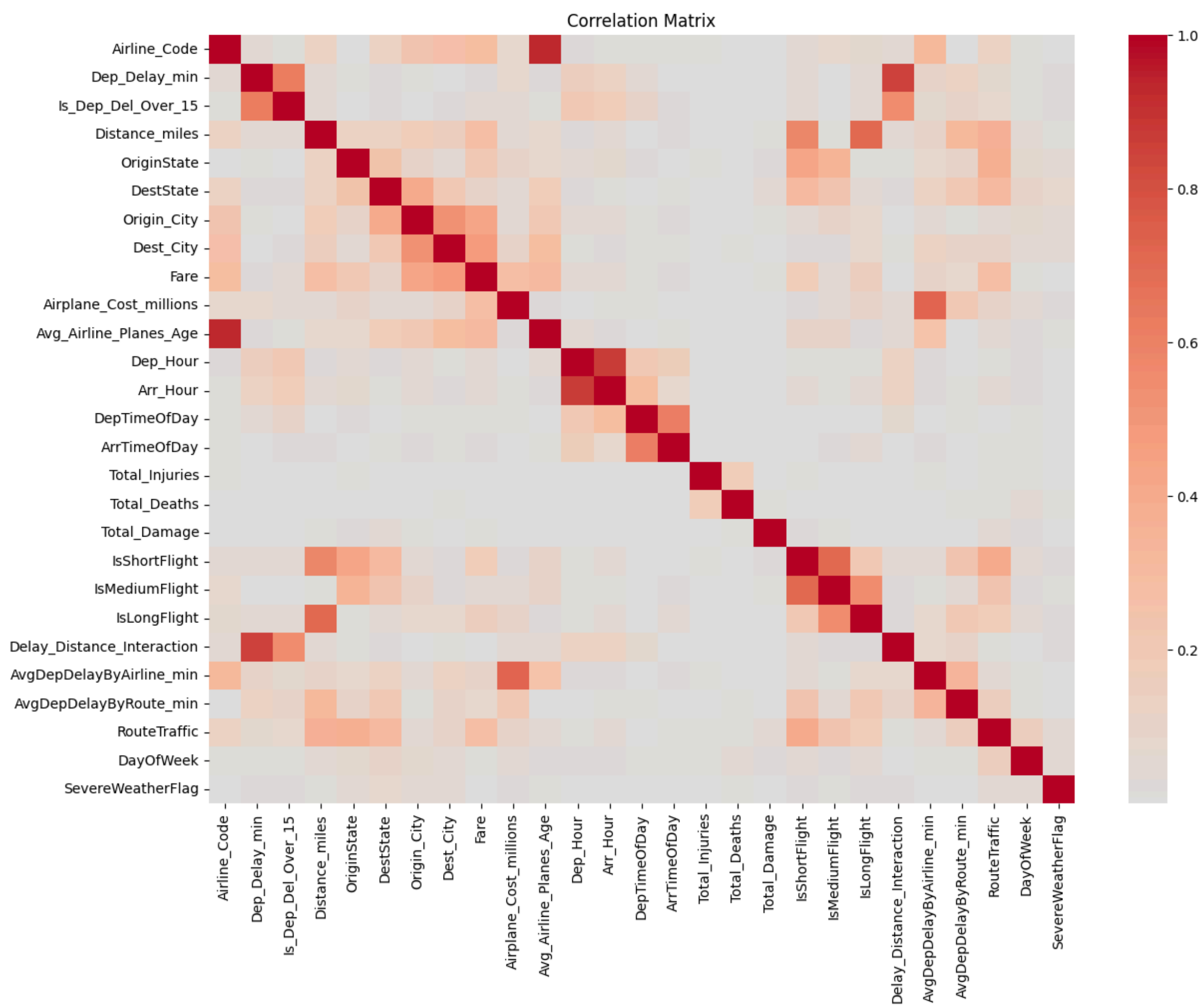
- **Time-Based Features** - We extracted scheduled departure and arrival hours and grouped them into broader time-of-day categories (morning, afternoon, evening, night). A new DayOfWeek feature was also added, capturing the effect of weekly patterns on delays.
- **Storm Impact Features** - To simplify damage-related data, we created Total_Injuries, Total_Deaths, and Total_Damage by combining direct and indirect counts. These features offer a more holistic look at event severity without requiring multiple columns.
- **Flight Traffic Features** - Flight distance was binned into categories (short, medium, long), with additional binary indicators for each. A Delay-Distance Interaction term was added to capture how distance and delay multiply together in terms of impact. We also created a Route feature combining origin and destination, allowing us to examine performance by specific routes.
- **Average delay** values by airline and route were calculated, helping highlight consistently underperforming airlines or routes. Another useful feature, RouteTraffic, was created to reflect how busy a route is on a specific day.
- **Weather-Based Features** - We introduced a SevereWeatherFlag by checking if any events with high magnitude occurred in the destination state on the flight date. This acts as a binary signal for potential weather disruptions.
- **Categorical Feature Encoding** - We encoded categorical variables based on their cardinality. Columns with a small number of unique values (like DepTimeOfDay) were One-Hot Encoded, while those with many values (like Airline_Code) were Label Encoded. This strategy helped maintain model performance without increasing dimensionality too much.

Feature Selection:

Target Variable: Is_Arr_Del_Over_15_min - indicates whether a flight's arrival was delayed by more than 15 minutes.

Correlation Analysis

We performed correlation analysis on numeric features and removed highly correlated (r > 0.85) variables to prevent multicollinearity and simplify the model.



Model Training and Evaluation:

Logistic Regression:

Pros: Fast to train, interpretable, simple baseline.
Cons: Struggles to capture complex, nonlinear patterns in the data, resulting in high false negatives — i.e., many delayed flights go undetected.
Use case: Good for quick prototyping or environments where explainability is more important than accuracy.

Random Forest:
Pros: Outstanding performance across all metrics — especially F1-score (0.9957) and AUC (0.9999). Very low false positives and false negatives.
Cons: May slightly overfit due to its ensemble depth; larger memory footprint.
Use case: Ideal for internal tools, backend services, or where predictive performance is the top priority.

XGBoost:
Pros: Strikes an excellent balance between performance and generalization. High F1-score (0.9338) and AUC (0.9941), with robust delay detection capabilities and faster inference than Random Forest.
Cons: Slightly lower recall than Random Forest, though still very strong.
Use case: Best suited for real-time dashboards, APIs, and production environments due to its speed and stability.

Confusion matrix:
Logistic Regression

	Predicted: 0 (On-Time)	Predicted: 1 (Delayed)
Actual: 0	537,804	20,909
Actual: 1	47,808	160,555

- False Negatives (missed delays): 47,808
- False Positives (falsely predicted delays): 20,909

Random Forest

	Predicted: 0 (On-Time)	Predicted: 1 (Delayed)
Actual: 0	558,074	639
Actual: 1	1,160	207,203

- False Negatives: 1,160
- False Positives: 639

XGBoost

	Predicted: 0 (On-Time)	Predicted: 1 (Delayed)
Actual: 0	549,185	9,528
Actual: 1	17,543	190,820

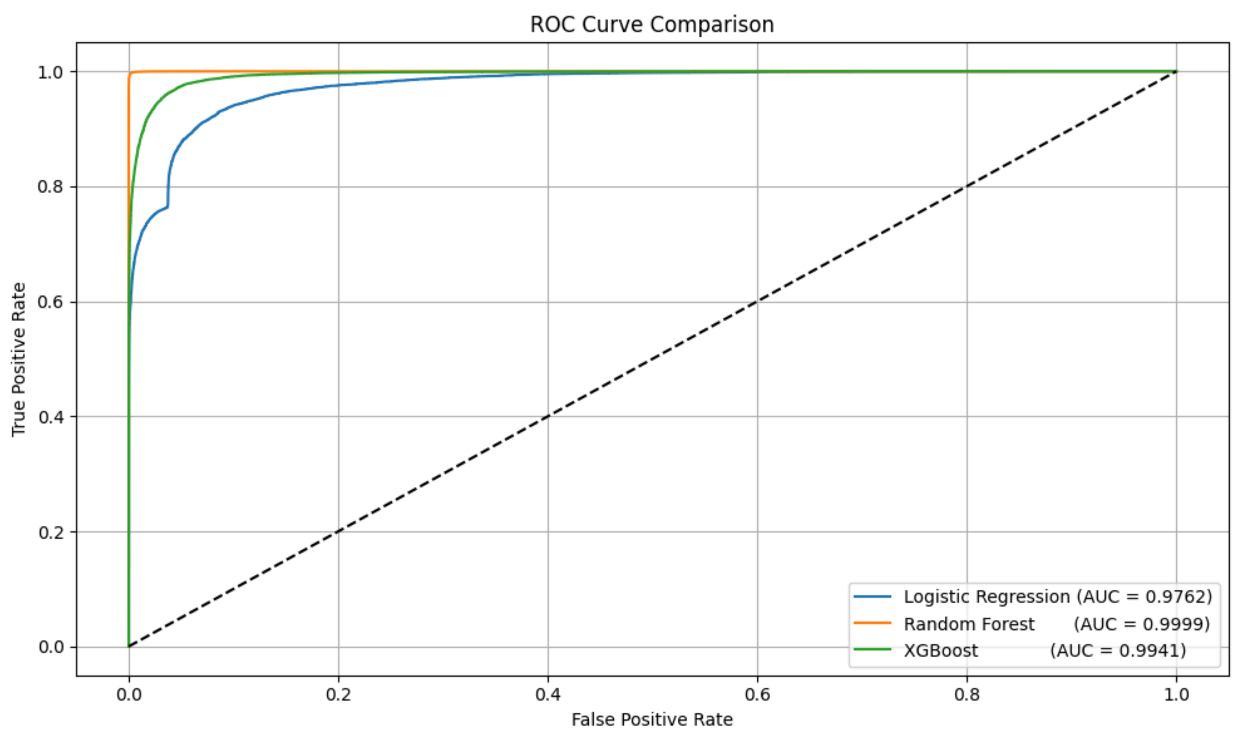
- False Negatives: 17,543
- False Positives: 9,528

Logistic Regression: High number of missed delays (false negatives), indicating underprediction of delayed flights.

Random Forest: Near-perfect classification with minimal false positives and false negatives.

XGBoost: Excellent performance with a good balance between detecting delays and avoiding false alarms.

ROC Curve:



ROC Curve Analysis:
All three models show strong ROC curves, with Random Forest achieving near-perfect separation between classes (AUC = 0.9999). XGBoost is close behind. Logistic Regression lags slightly, especially at lower false positive rates.

Deep Learning Approach: Fully Connected Neural Network (FCNN)

We also experimented with a Fully Connected Neural Network (FCNN) to see how deep learning would perform on the flight delay classification task.

Model Setup - The model had an input layer that matched the number of features, followed by three hidden layers using ReLU activation and dropout to prevent overfitting. The output layer used a sigmoid function to predict whether a flight would be delayed or not. We used binary cross-entropy as the loss function and the Adam optimizer for training.

Training and Tuning - We added early stopping and learning rate scheduling to improve generalization. Hyperparameter tuning was done using KerasTuner and took about 52 minutes. The best model reached:

- **Validation MAE**: 0.0640
- **Training Loss**: 0.0299
- **Training MAE**: 0.0642

Conclusion:

Although the FCNN gave decent results, it didn't beat models like Random Forest or XGBoost. Neural networks can be powerful, especially at scale with GPUs, but for this project, the simpler tree-based models performed better and were easier to interpret.

Interpretation of model outputs:

Dashboard Overview & Why It Works

As part of this project, we built an interactive **Streamlit dashboard** for flight delay prediction. This dashboard makes it easy to input flight details, generate predictions, and understand patterns from past predictions.

A dashboard is ideal for our use case because:

- It provides a **simple and intuitive UI** for non-technical users
- It combines **real-time prediction** with **historical insights**
- It allows users to **interact** with the model, not just view static outputs

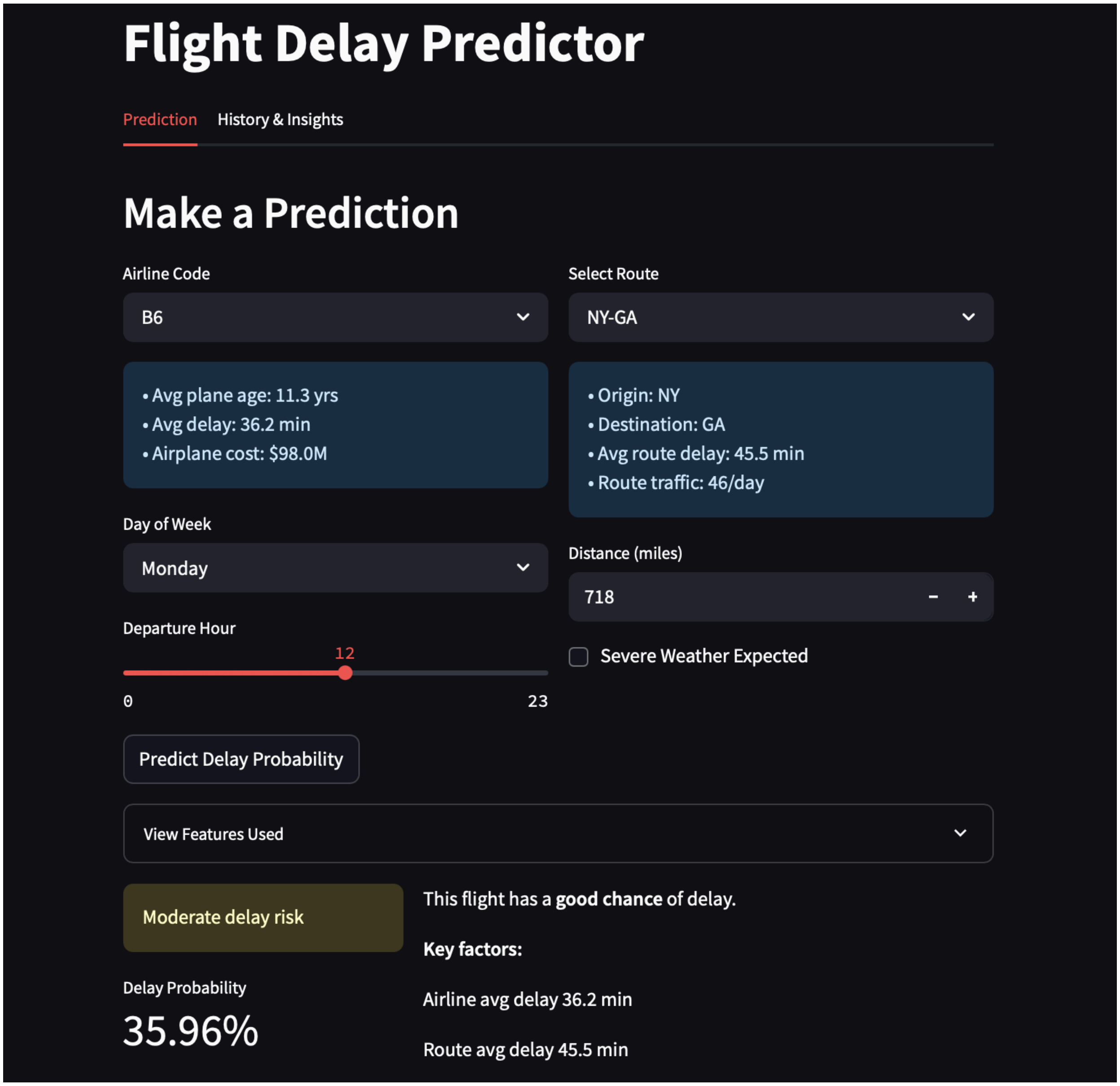
Two Tabs: Prediction and History & Insights

Tab 1: Prediction

In the first tab, users can:

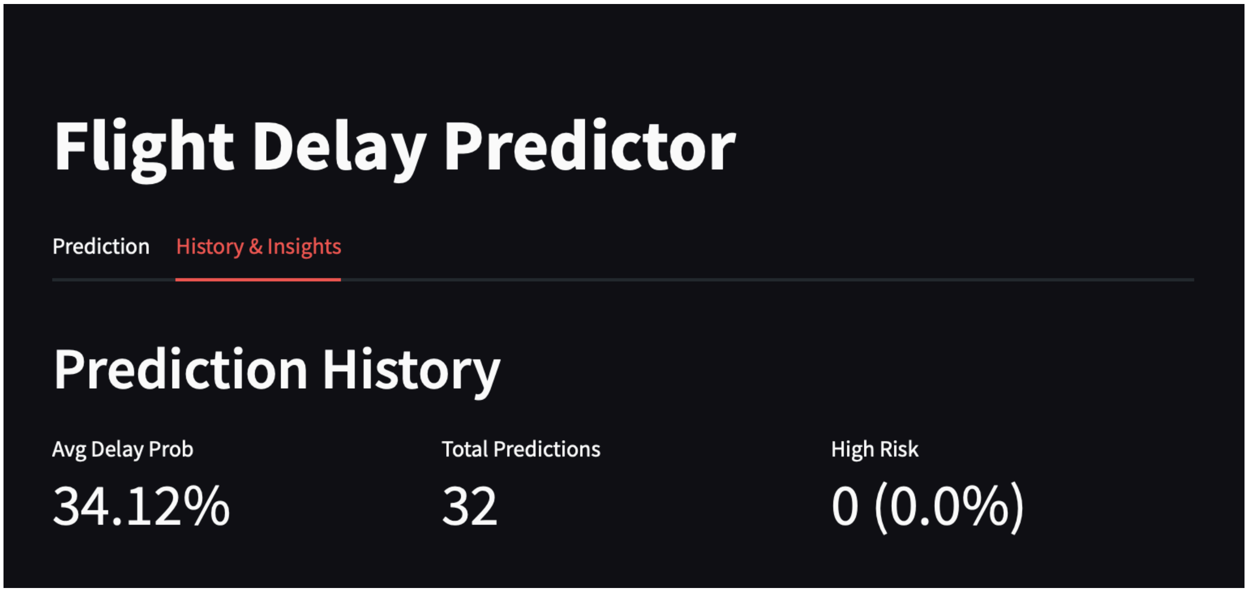
- **Select an airline code**
- **Choose a route** (origin and destination states)
- See relevant info like:
  - Average plane age
  - Average delay time
  - Airplane cost for that airline
- **Pick a day of the week** (Monday to Sunday)
- **Input the distance** of the flight (in miles)
- **Toggle severe weather** if expected
- **Select the hour of departure** using a slider

After setting these inputs, clicking **“Predict Delay Probability”** runs the model and shows the predicted risk, along with the key factors that contributed to the prediction.

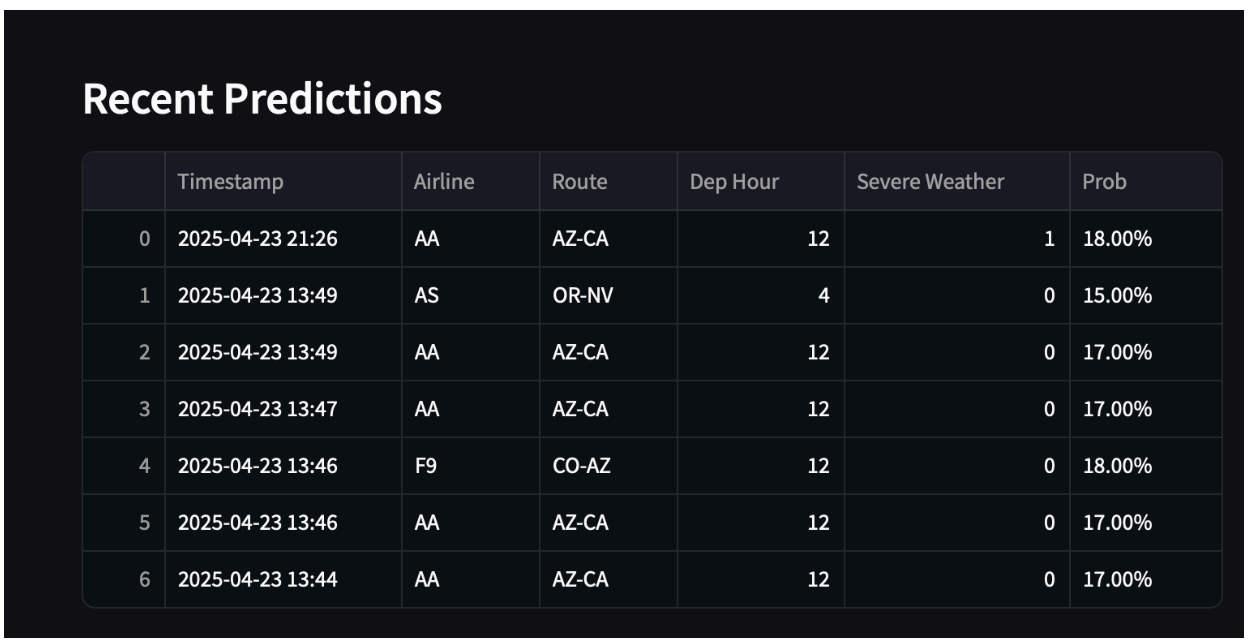


Tab 2: History & Insights

In the second tab, we show insights from past predictions. Every prediction made through the dashboard is saved to a local **SQLite database**, and the following is displayed:



- **Average delay probability** across all predictions
- **Total number of predictions** made
- **Count of high-risk predictions** (delay probability > 60%)
- A **table of recent predictions** with timestamp, airline, route, dep hour, weather, and predicted delay %

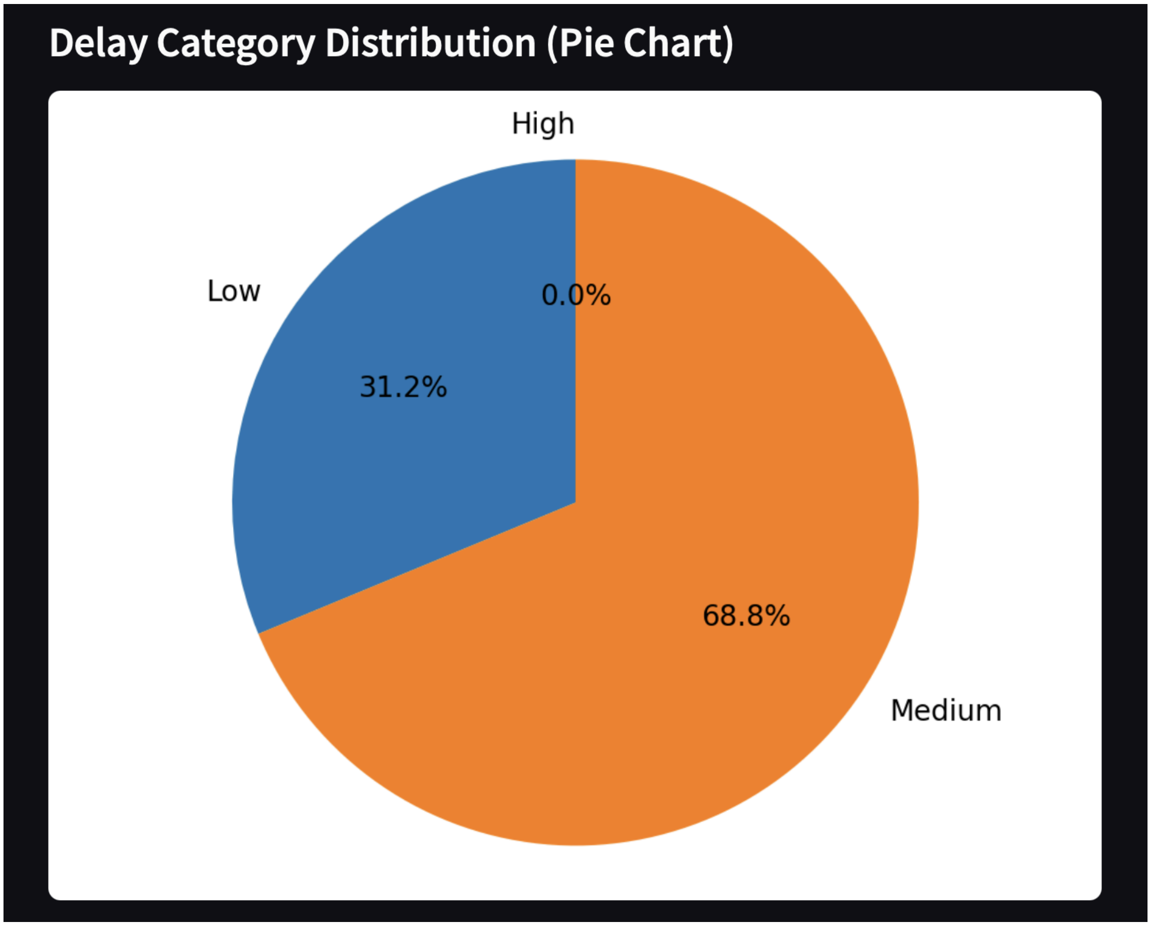


Recent Predictions

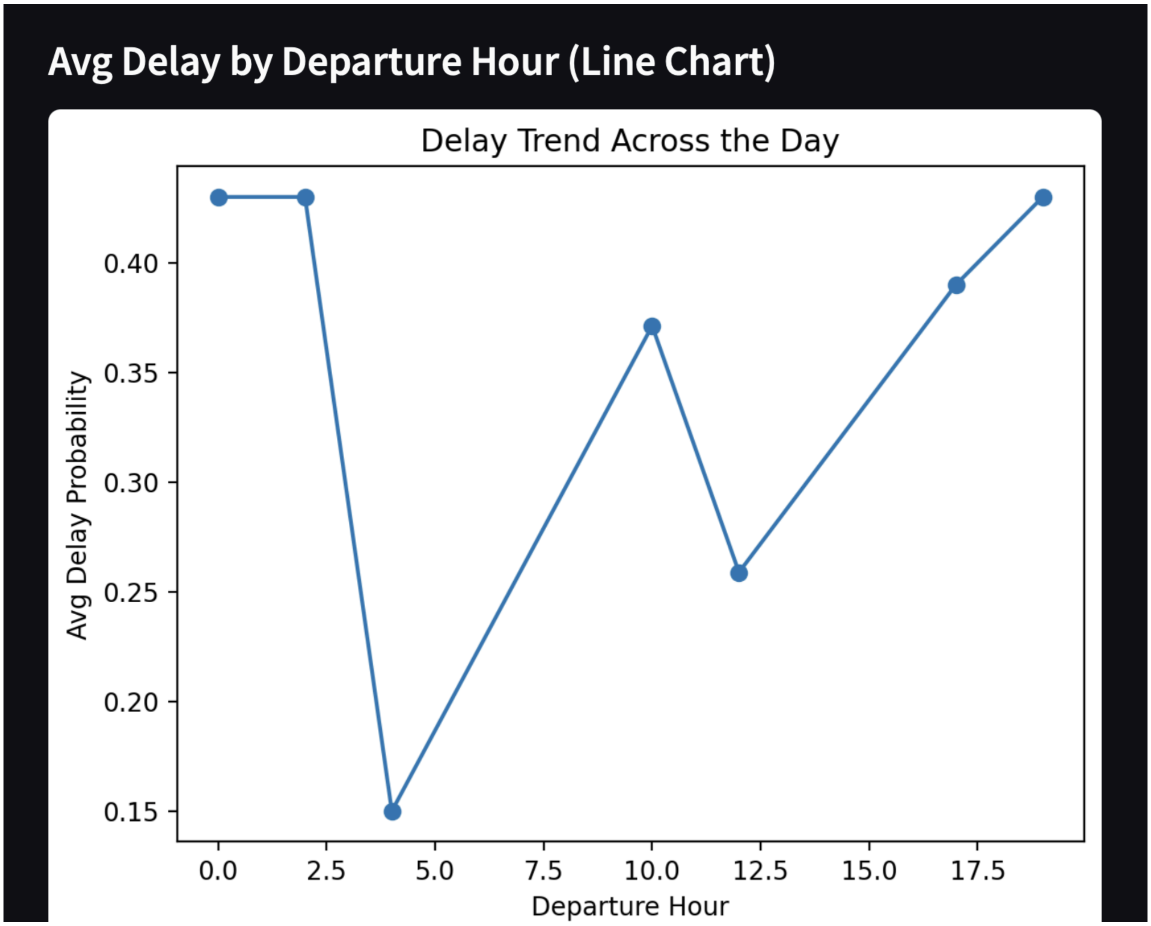
	Timestamp	Airline	Route	Dep Hour	Severe Weather	Prob
0	2025-04-23 21:26	AA	AZ-CA	12	1	18.00%
1	2025-04-23 13:49	AS	OR-NV	4	0	15.00%
2	2025-04-23 13:49	AA	AZ-CA	12	0	17.00%
3	2025-04-23 13:47	AA	AZ-CA	12	0	17.00%
4	2025-04-23 13:46	F9	CO-AZ	12	0	18.00%
5	2025-04-23 13:46	AA	AZ-CA	12	0	17.00%
6	2025-04-23 13:44	AA	AZ-CA	12	0	17.00%

Insights:

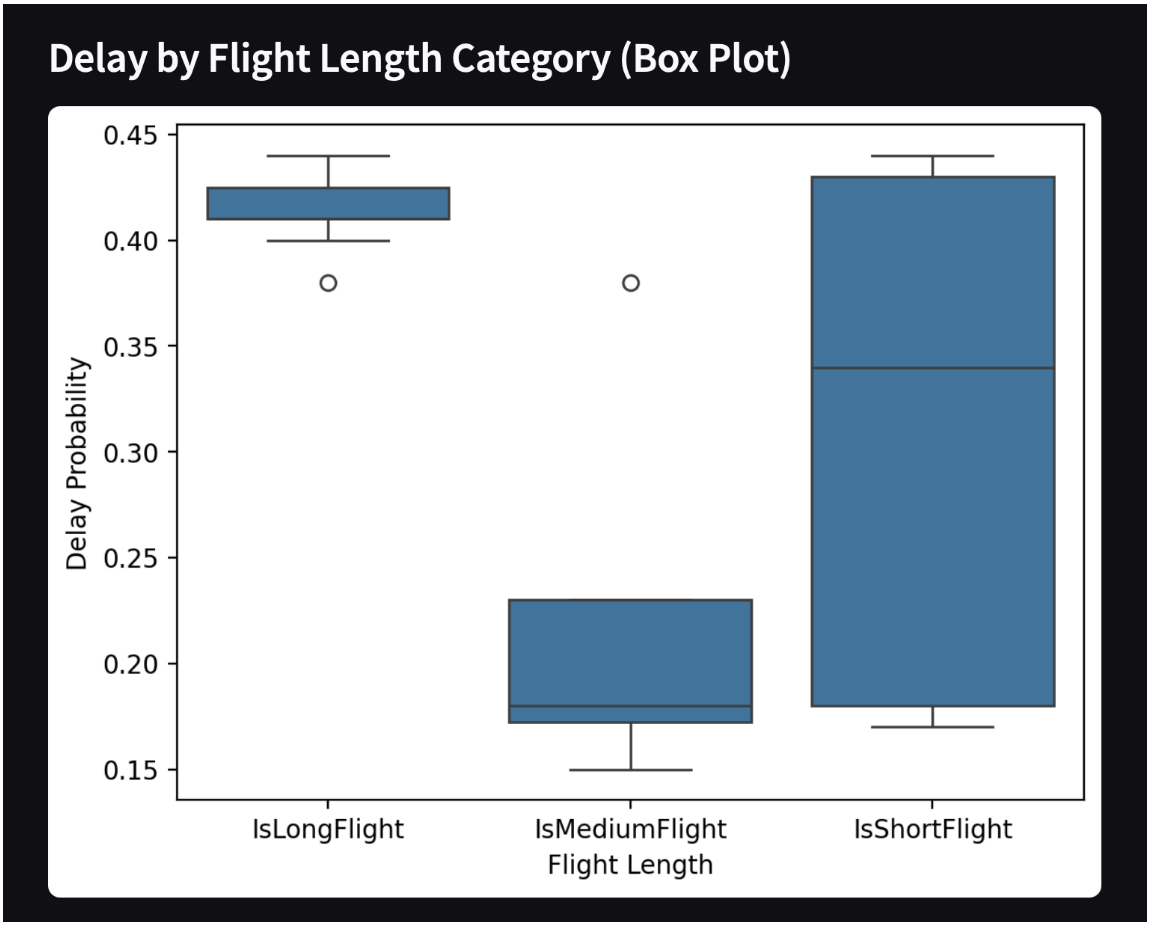
1. **Pie chart of delay categories**: Shows the proportion of predictions classified as Low, Medium, and High delay risk.



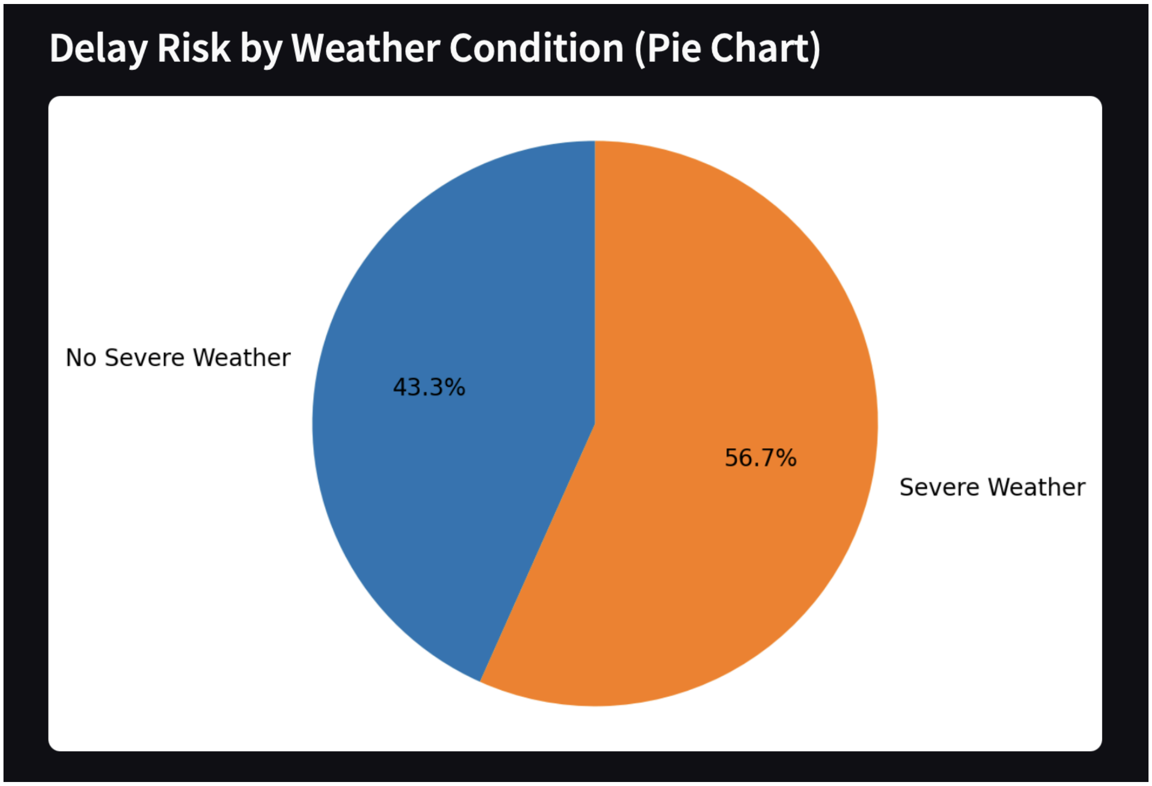
2. **Line chart of average delay by departure hour**: Reveals time-of-day trends in flight delays.



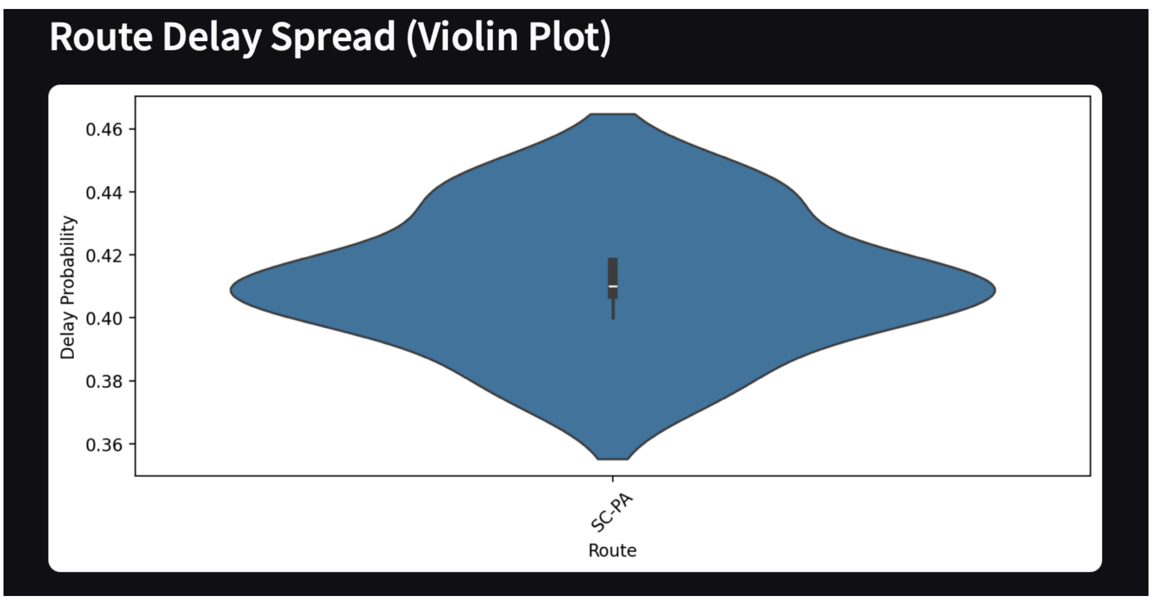
3. **Box plot of delay probability by flight length**: Compares delay risk for short, medium, and long flights.



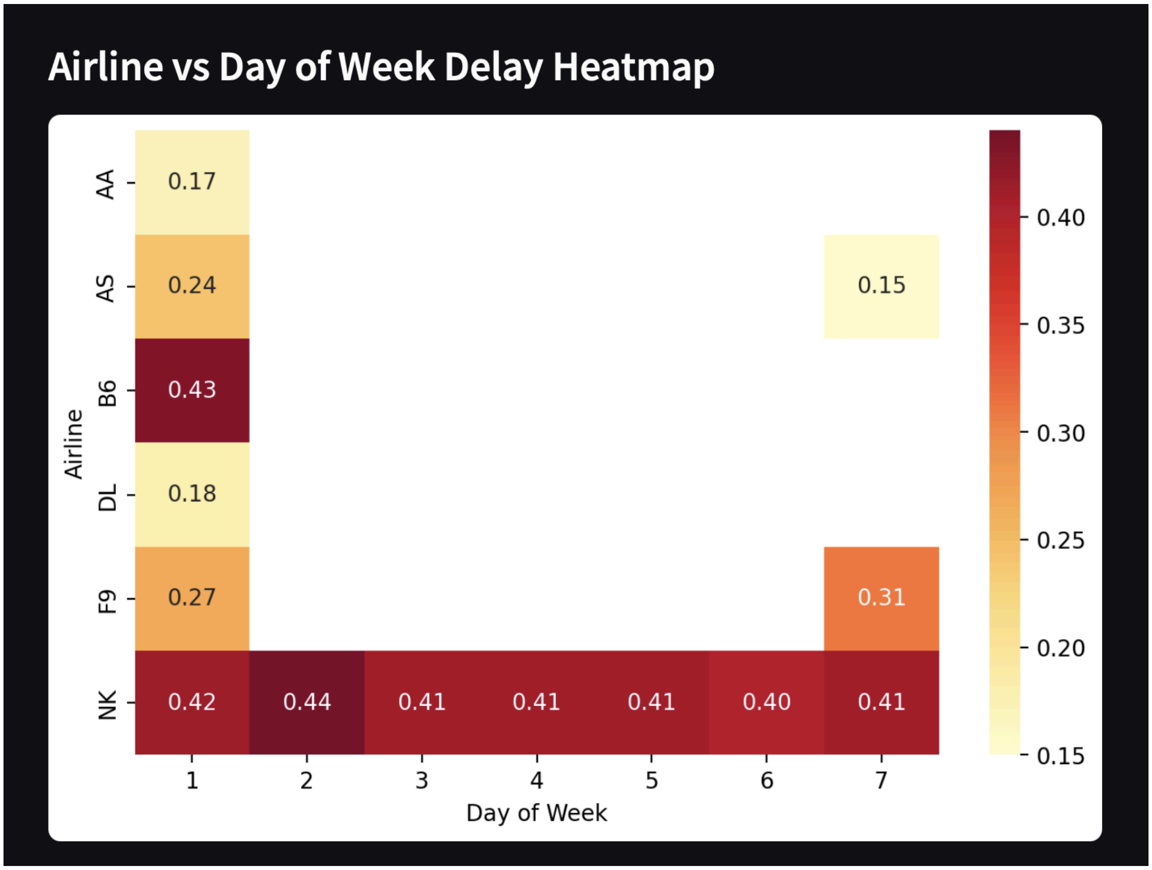
4. **Pie chart of average delay by weather condition**: Highlights how severe weather affects delay probabilities.



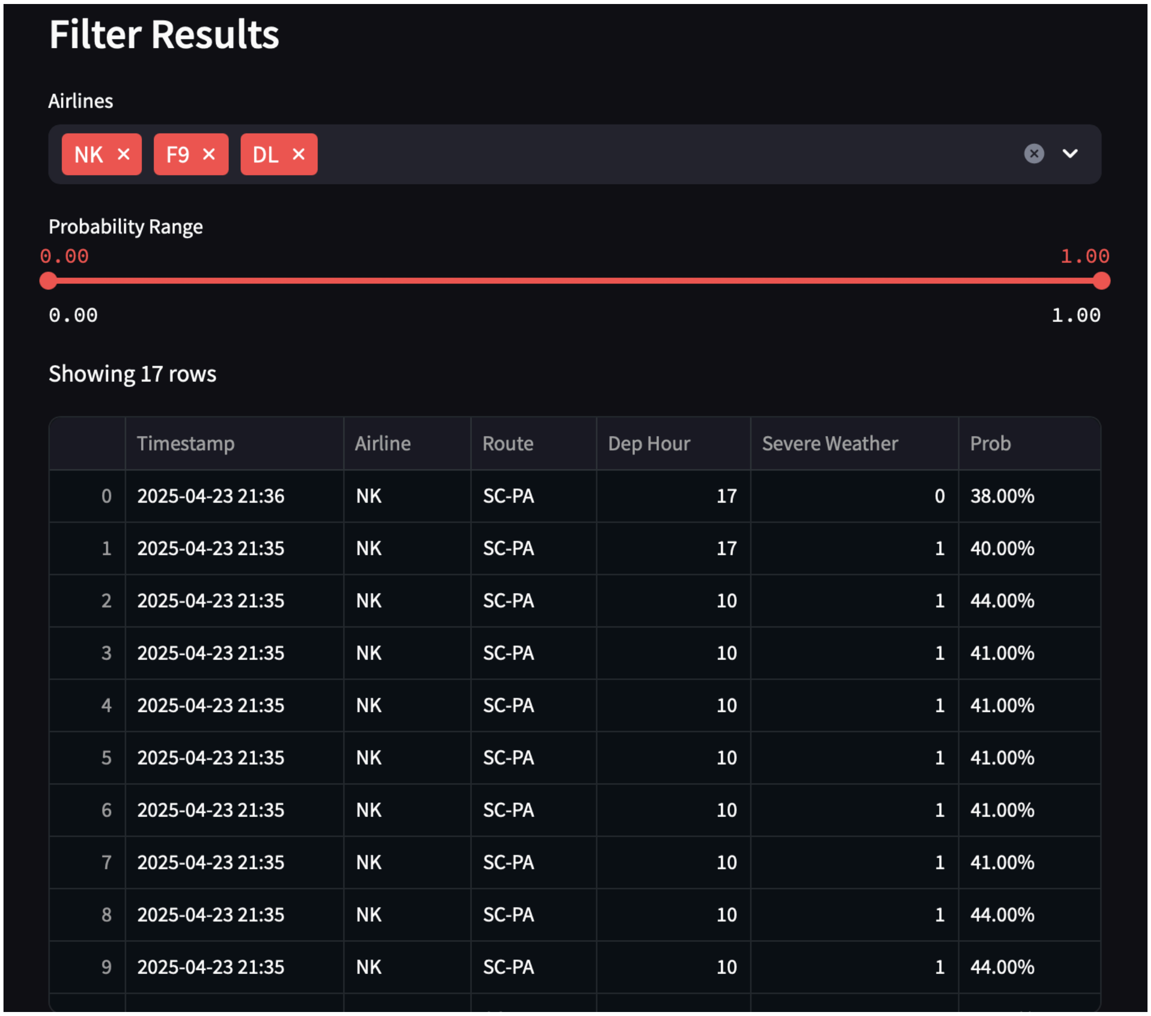
5. **Violin plot of route-level delay spread**: Shows how delay probabilities vary across different routes.



6. **Heatmap of delay risk by airline and day of week**: Helps identify which airlines perform better on certain days.



7. **Filtered view**: Lets users filter predictions by airline and delay probability range.



Filter Results

Airlines

NK × F9 × DL ×

Probability Range

0.00 1.00

0.00 1.00

Showing 17 rows

	Timestamp	Airline	Route	Dep Hour	Severe Weather	Prob
0	2025-04-23 21:36	NK	SC-PA	17	0	38.00%
1	2025-04-23 21:35	NK	SC-PA	17	1	40.00%
2	2025-04-23 21:35	NK	SC-PA	10	1	44.00%
3	2025-04-23 21:35	NK	SC-PA	10	1	41.00%
4	2025-04-23 21:35	NK	SC-PA	10	1	41.00%
5	2025-04-23 21:35	NK	SC-PA	10	1	41.00%
6	2025-04-23 21:35	NK	SC-PA	10	1	41.00%
7	2025-04-23 21:35	NK	SC-PA	10	1	41.00%
8	2025-04-23 21:35	NK	SC-PA	10	1	44.00%
9	2025-04-23 21:35	NK	SC-PA	10	1	44.00%

These help users understand not just individual predictions, but also **trends** like which airlines are more reliable or which routes are delay-prone.

Limitations and Biases:

1. **Manual Severe Weather Flag** - We use a checkbox input instead of real-time weather data, which may not reflect actual weather conditions accurately during prediction.
2. **No Real-Time Flight Update** - Flight delays often depend on live operational data (e.g., gate availability, previous flight delays), which our model doesn't access.
3. **Limited Data Scope -** The training data is filtered to a single year (2020), which may not generalize well across different years or unusual travel trends.
4. **Route History Bias -** Delay patterns for a route depend heavily on past performance. New or rarely flown routes may be misrepresented or underrepresented.

Final Conclusion:

This project achieved its goal of building an interactive dashboard that predicts flight delays using machine learning and real-world data. We combined multiple datasets—flights, weather, airline info, and fares—to understand what really drives delays.

The XGBoost model, paired with PCA, performed well, but the real value lies in the dashboard. Users can input flight details and instantly get delay predictions, along with easy-to-understand reasons like weather, airline history, or time of day.

The insights tab adds even more value by showing patterns like which airlines or routes are riskier, and how timing affects delays. While we don't use real-time weather data, the system still helps users and airlines make more informed decisions.

In short, this project blends accurate predictions with clear explanations—making flight delays easier to understand and plan for.