Flight Delay and Cancellation Analysis (2019-2023)

Bishal Rauniyar Devansh Trivedi Manish Thapa

1 GitHub

https://github.com/Thapamanish/flight-delay-prediction

2 Abstract

Delays and cancellations of flights have been an ongoing source of cost and frustration for aviation in the U.S. In this project, we analyze a large-scale dataset comprising about 3 million domestic flights operated between January 2019 and August 2023. It covers 18 airlines and approximately 380 origin and destination airports.

Our flow is database-first: all preprocessing, feature engineering, and exploratory statistics reside as SQL queries running on DuckDB, with Python used mostly for visualization and modeling. We have performed exploratory data analysis to see how delays vary by carrier, time, airport, and cause. Thereafter, two classification scenarios are developed using logistic regression and Random Forest models. The pre-flight model, i.e., based only on information known at booking time, and the real-time model that also uses actual departure delay.

In our observation, the Random Forest gets an AUC of 0.653, but it gives actionable risk estimates. Real-time gets an AUC of 0.956 with 91% precision and 77% recall. This means the departure delay is the dominant driver of arrival delay. Feature-importance analysis shares with us that about 84% of the predictive power of this model comes from departure delay which is consistent with our EDA finding that late aircraft cascading is the primary delay cause.

3 Introduction

Millions of passengers are affected by flight delays and cancellations, creating huge operational and financial challenges to the airlines and airports. These small disruptions typically snowball into a tightly coupled schedule system, increasing impacts from weather events, congestion, and maintenance issues. Therefore, knowing where, when, and why delays happen is fundamental toward any effort aimed at improved reliability and enhanced passenger experience.

1. We've describe patterns of flight delays and cancellations across airlines, airports, and time periods, including the impact of the COVID-19 pandemic.

2. Quantify the contribution of different delay causes (late aircraft, carrier, weather, National Airspace System (NAS), security).

3. Have build and evaluate classification models to predict whether a flight will arrive late, under two scenarios:

   3.1. Pre-Flight Prediction: only booking-time information is available.

   3.2. Real-Time Prediction: actual departure delay is also known.

We do not use pandas as our preferred engine for analysis, but rather apply a database-centric design. DuckDB offers in-process SQL analytics scaling up to millions of rows, maintaining the same expressiveness and reproducibility that SQL provides.

4 Data and Infrastructure

4.1 Dataset Overview

The dataset we have used is from the U.S. Department of Transportation's Bureau of Transportation Statistics (On-Time Performance data, accessed via Kaggle). It covers domestic flights within the United States from January 2019 to August 2023. The sample contains approximately three million flight records with 32 attributes, representing about 18 airlines and nearly 380 origin and destination airports. Each record includes flight details, operational status, delay metrics, cancellation information, and scheduled as well as actual departure and arrival times.

Key attributes
The data we used contains details at the flight level including date of flight, airline, flight number and origin-destination airports. It also captures the status of operation for each flight (performed, delayed, cancelled, or diverted) along with arrival and departure delay value as well as minutes related to delays by cause (carrier, weather, NAS constraints, security events, and late aircraft). The cancellation codes indicate the primary reason for any cancelled flight. This dataset includes scheduled Real departure and arrival times help make time-focused features.

Simple descriptive statistics from the unprocessed table indicate that these 3 million flights cover 18 airlines and approximately 380 unique origin as well as destination airports. The mean value for great-circle distance is about 809 miles while the average arrival delay for all flights (including those arriving early) stands at 4.26 minutes.

4.2 Database-First Workflow with DuckDB

All core data operations are implemented in SQL using DuckDB:

So we have loaded the raw CSVs into a DuckDB table called flights_raw, run data quality and basic summary stats right there in SQL. Create a cleaned dataset version named flights_clean by filtering and validating records-plus extracting year month day of week fields-also adding distance and duration fields as well as creating a binary label IS_DELAYED field for any flight arriving ten or more minutes late. Also create another table delay_reasons to be used for aggregating the specific reason codes (carrier, weather, NAS, security, plus late aircraft) delay minutes.

Python acts mostly as a client: it runs SQL through DuckDB, gets results converted to DataFrames, and visualizations outputted by Matplotlib/Seaborn. This is the best way to reflect real production data-engineering patterns while also keeping the notebook tight and reproducible.

5 Data Quality and Preprocessing

5.1 Missing Data Assessment

We started by checking the level of missingness in some important columns (such as arrival delay, delay causes, cancellation codes, and airline identifiers).

- Cause-specific delay fields (DELAY_DUE_CARRIER, DELAY_DUE_WEATHER, DELAY_DUE_NAS) and CANCELLATION_CODE have high missingness (≈80-97%), because they are populated only when a flight is delayed or cancelled.
- Core fields such as airline and flight date have no missing values.

This pattern is expected given the data collection process, we therefore treat cause fields as conditionally observed rather than globally missing.

5.2 Dataset Summary and Cleaning

Using SQL:

We compute several descriptive statistics for the dataset, including the total number of flights, the number of unique airlines and airports represented, the average great-circle distance and mean arrival delay, and the counts of cancelled and diverted flights.

We then create a cleaned dataset flights_clean with the following steps:

1. **Filter to valid records**
   - Keep domestic, non-null flights with valid time and distance fields.
2. **Create derived columns**
   - Year, month, and day-of-week from flight date
   - Distance bands (e.g., <500 miles, 500-1000, 1000-1500, >1500)
   - IS_DELAYED = 1 if arrival delay $\geq$ 10 minutes, else 0.
3. **Prepare for classification**
   - Categorical features (airline, origin, destination, time buckets) are kept as codes to be one-hot encoded later in the modeling notebook.

6 Exploratory Data Analysis

6.1 Airline and Destination Composition

We first examine which airlines and airports dominate the sample.

A bar chart of the top 15 airlines by flight count shows that Southwest, Delta, American, SkyWest, and United account for a large share of total flights, with Southwest alone operating nearly one-fifth of all flights in the sample. A scatter chart of top destination airports (by total flights and delay rate) shows major hubs such as ATL, DEN, LAX, and ORD. Some airports combine high traffic with above-average delay rates, making them critical nodes for network reliability.

Delays at a small number of large carriers and hub airports have disproportionate impact, because they affect a large fraction of travelers and aircraft rotations.

6.2 Day-of-Week Patterns

Using SQL, we have computed for each day of the week:

- total flights,
- number of delayed flights,
- delay rate (% of flights delayed),
- average delay minutes among all flights.

Results show:

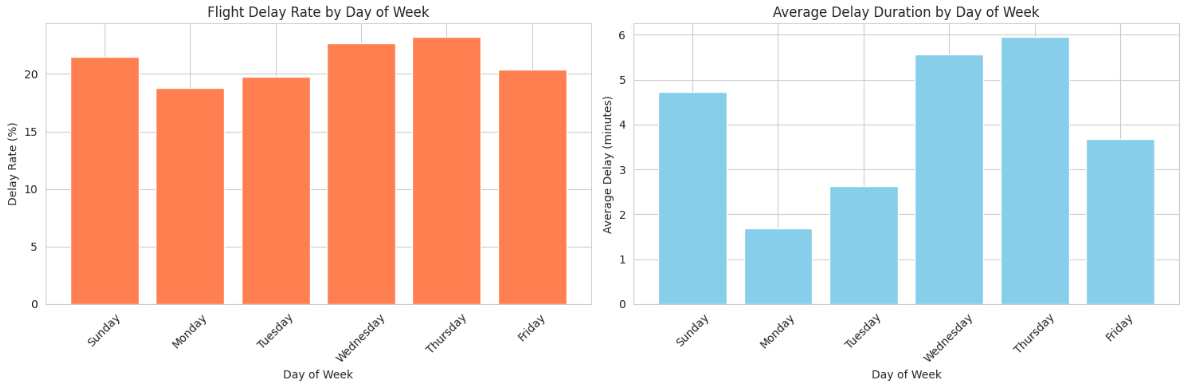

Figure 1: Delay rate and average delay by day of the week.

- Delay rates range from about 19-23% across weekdays.
- Wednesday and Thursday shows the highest delay rates (∼22-23%) and the longest average delay durations (∼5.5-6 minutes).
- Monday has relatively lower delay rates and shorter average delays.

Mid-week congestion and tightly packed schedules may contribute to higher delays, while Mondays benefit from “reset” schedules and lower residual cascading.

6.3 Annual and Monthly Trends (COVID-19 Impact)

Yearly aggregation reveals:
In 2019, about 22% of flights were delayed, with total delay minutes around 3.9 million.In 2020, during the peak of COVID-19 restrictions, traffic collapsed (≈450K flights vs 740K in 2019), delay rates dropped to about 12%, and net arrival delays became negative on average (flights tended to arrive early). By 2022-2023, traffic and delays not only recovered but exceeded 2019 levels, with

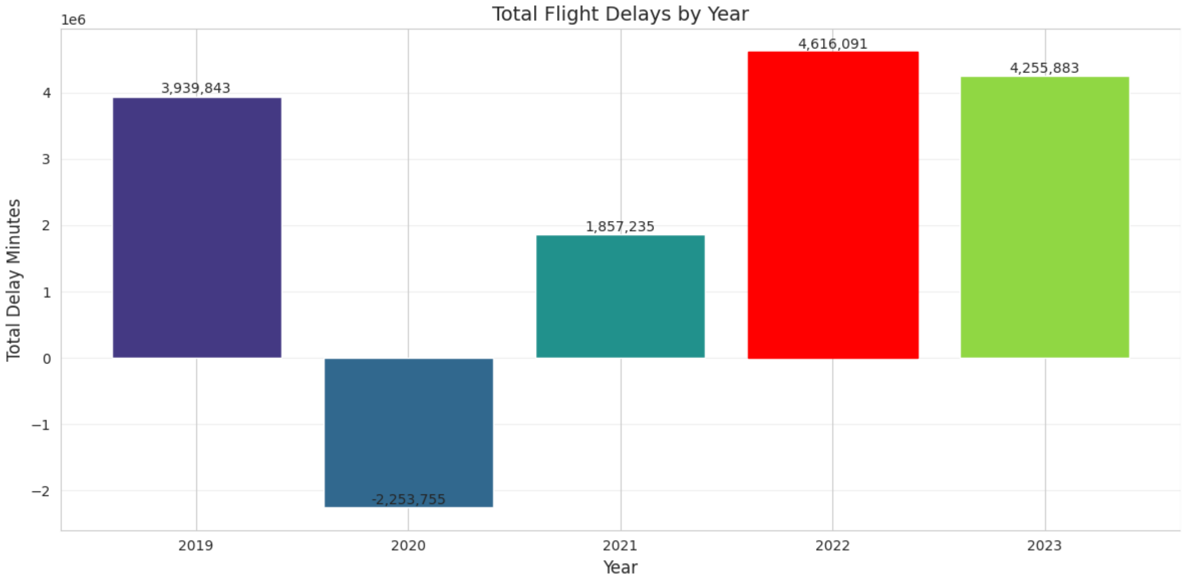

Figure 2: Total annual delay minutes, highlighting the 2020 dip and the 2022 peak.

total delay minutes peaking at roughly 4.6 million minutes in 2022 and high delay rates continuing into 2023.

Monthly trend plots also illustrate the strong summer peak and relatively quiet winter months, with 2020 being an exceptionally low-delay year. The pandemic dramatically reduced both volume and delays in 2020. As demand rebounded faster than infrastructure and staffing in 2022-2023, congestion and delays increased beyond pre-pandemic levels.

6.4 Delay Causes

Aggregating cause-specific delay minutes across the dataset, we obtain the breakdown shown in the figure below:

Late aircraft and carrier related minutes dominate the dataset with 13.6 million minutes (37.7%) and 13.2 million minutes (36.7%) respectively, NAS related delays added another 7.0 million minutes (19.5%), weather delays accounted for 2.1 million minutes (5.9%). Security related delays are insignificant, only about 0.08 million minutes (0.2%).

Carrier and late-aircraft delays plummeted in 2020, rose again in 2021, reached a peak in 2022, and slightly improved in 2023.These were the two large components of total delay. Other smaller components, NAS and weather, followed similar albeit smaller magnitude patterns.
Late aircraft and carrier operational issues summed up almost three-fourths of the total delay minutes. Since late aircraft delays come from previous flights, it serves as a very good indicator of cascading disruptions which will later on prove to be the most important feature in our predictive models.

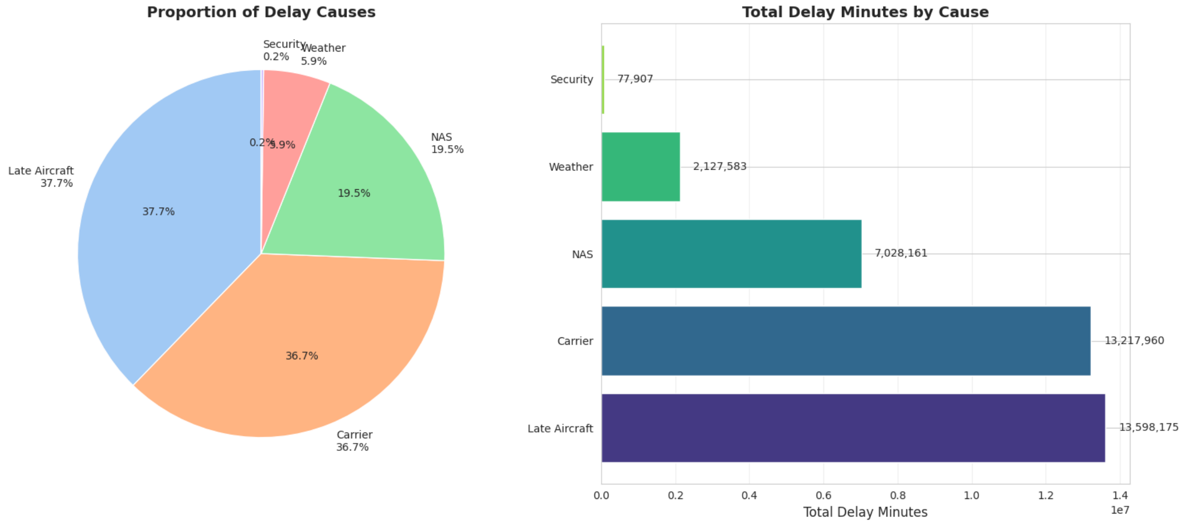

Figure 3: Visualizes the share of each delay cause, with late aircraft and carrier-related delays dominating.

6.5 Routes and Distance

Advanced SQL queries identify:
The most problematic routes by delay rate are the top 20. Major hubs or weather-sensitive airports have many of them. There is a bar chart of delay rate by distance band. Medium to long haul flights have slightly higher delay rates than very short flights. The differences though, are modest.

Some origin-destination pairs just keep being unreliable, so adding some extra buffer time or adjusting the schedule might really help.

7 Predictive Modeling

The EDA shows strong patterns about delays in association with late aircraft departure or late aircraft arrival. To measure how predictable delays are, we set up two classification scenarios in a different notebook for modeling inside Google Colab.

7.1 Problem Definition and Features

We define a binary label:

$$Y = 1 \text{ if arrival delay } \geq 15 \text{ minutes, and } Y = 0 \text{ otherwise.}$$

Two feature sets are considered:

In **Scenario A (Pre-Flight Prediction)**, we have used data known at or before the booking time. That shall include the carrier, origin–destination airport pair, scheduled departure time (binned into hour-of-day, day of week, month or season), and distance of the flight.

Scenario B (Real-Time Prediction) The feature set is extended by the actual minutes of departure delay (or optionally in coarse bins such as 0–5, 5–10, or $¿10$ minutes). This real-time operational variable presents a much stronger signal and therein lies the magic of dramatic improvements in predictive performance.

Data is split into training and test sets (e.g., 80/20). Categorical variables are one-hot encoded, continuous variables are standardized where appropriate.

7.2 Models and Evaluation Metrics

We use two families of models in scikit-learn: Logistic Regression, as the simple linear baseline, and Random Forests, being the main nonlinear model in both pre-flight and real-time prediction scenarios.

Some of the standard evaluation metrics that we used to judge this model included AUC, Accuracy, Precision, Recall, and F1-score. These metrics provide a good view not only regarding how well these models can discriminate but also about their performance in terms of balance between false positives and false negatives.

7.3 Scenario A Pre-Flight Prediction

The booking-time features only result in an AUC of 0.653 from the Random Forest on the held-out test set, which is approximately 30% better than random guessing (AUC 0.5). It carries a moderate overall accuracy and F1-score because most delays are driven by factors that occur during the day-of-operations, information not available at the time of booking.

Logistic regression performs slightly worse than Random Forest, thereby confirming that there do exist some nonlinear interactions between airline, route, and schedule. However, it should be noted that flights from both these models present probabilistic risk scores which can readily be used in ranking flights by their expected delay risk.

The pre-flight Random Forest achieves a ROC AUC of approximately 0.65, which confirms that using only pre-flight information provides limited ability to separate delayed flights from on-time flights.

As shown in Figure 4 above, the real-time model correctly captures most delayed flights while keeping a low false-positive rate.

Even when fed with little information, if it can beat random odds at distinguishing between low-risk and high-risk flights, then perhaps this is the data that passengers need to select less risky itineraries, or airlines want to audit a route when a model continually flags it with high predicted delay probability.

7.4 Scenario B Real-Time Prediction

When we add departure delay as a feature, model performance improves dramatically:

- The Random Forest reaches an AUC of 0.956.
- On the test set it achieves 91% precision and 77% recall for predicting delayed arrivals.

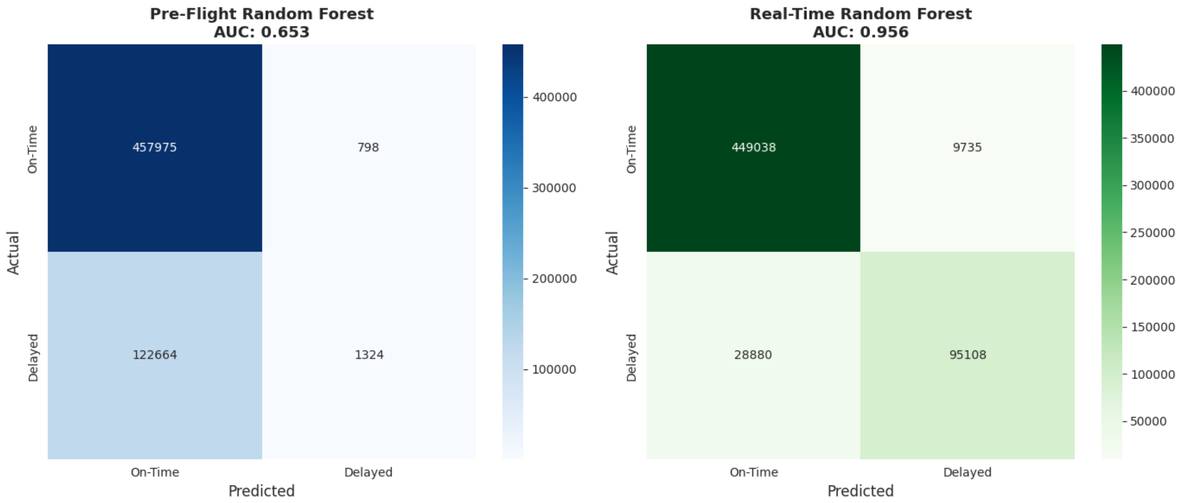

Figure 4: Confusion matrices for the pre-flight (left) and real-time (right) Random Forest models. The titles show the ROC AUC for each setting, illustrating the large gain in discriminative power once real-time departure delay is available.

A confusion-matrix plot shows that the model has been able to pick up the overwhelming majority of delayed flights with relatively low false alarm rates.

In contrast, the real-time Random Forest reaches a ROC AUC of about 0.96, indicating that once departure delay is observed the model can almost perfectly distinguish delayed vs. on-time arrivals. This large jump from roughly 0.65 to 0.96 AUC highlights how powerful real-time information is compared to purely pre-flight features.

Once a flight is late departing, knowing how late it was at departure pretty much fully determines whether it will arrive late. The about 30-point AUC improvement from adding this single feature quantifies cascading effects of late departures on arrival delays and validates our EDA finding that late aircraft delay is the dominant cause.

7.5 Feature Importance

A feature-importance analysis of the Random Forest in Scenario B shows that:

- Departure delay alone accounts for about 84% of the model's total importance score.
- The remaining signal is distributed across airline, route, time-of-day, and day-of-week features.

This proves that operational condition at departure causes arrival delays: if a flight leaves the gate in way late, then most probably it will also arrive late.
Operational implication: Interventions that may prevent or reduce early-day departure delays- that may include ground-time buffers, and improved turnaround processes as well as the prioritization of on-time departures for certain aircraft- would have multiplicative benefits by preventing delay cascades throughout the network.

8 Discussion

Some major patterns in the data are emphasized here. First, since a relatively small set of airlines, airports, and routes comprises abundant shares of total traffic and delays, this highly suggests that targeted improvements in these areas can have an oversized effect across the whole system. Second, delay behavior was radically transformed during the pandemic period. In 2020, volumes were about half as large as in 2019 and delays were much less prevalent, average arrival times became net slightly early. When demand returned in 2022–2023 delays not only returned but came back even worse than pre-pandemic levels indicated staffing and infrastructure capacity not keeping up with surging passenger volume.

Late aircraft and carrier-related issues make up about three-quarters of total delay minutes, proving that departure delay is the main reason for late arrivals. The level of predictability of delays largely depends on available information. With only booking-time features (Scenario A), the model has low discrimination with an AUC of 0.653. When actual departure delay is included (Scenario B), separability becomes nearly perfect with an AUC of 0.956. This comparison highlights just how valuable real-time operational signals are in forecasting arrival performance.

9 Conclusion and Future Work

9.1 Conclusion

This project puts together a big sample of three million U.S. inside-the-country flights from 2019 to 2023 using a database- first DuckDB and SQL workflow. Wide exploratory data analysis was carried out which showed the delay patterns by airlines, airports, time periods, and causes (reflecting the effects caused by the COVID-19 pandemic). Here it tries to quantify how late aircraft as well as carrier-related delays dominate total delay minutes plus highlighting several routes and distance bands particularly prone to disruptions.

We built and estimated pre-flight and real-time classification models using logistic regression and Random Forest, results indicated that booking-time information provides modest predictive power and that the actual departure delay allows near-perfect discrimination of delayed arrivals. In sum, this is how both structural factors airlines, routes, causes operate together with real-time operational conditions departure delays to shape arrival performance.

9.2 Future Work

This analysis opens up several directions for our future work. External data might well include, for example, more granular weather observations or information about restrictions by air-traffic control, which could explain more variance than features that can be included at the scheduling or carrier level. Possible approaches to better modeling include gradient boosting and XGBoost, and probability models that can be calibrated on real data, in this paper logistic regression is compared to Random Forest as a baseline method. Further possible developments involve creating an interactive dashboard application that would supply users with delay risk scores calculated using Scenario B in near real time. What-if analyses quantifying how such measures as interventions to reduce departure delay early in the day or addition of buffer time on specific routes would impact congestion downstream in the network would be very informative from an operational perspective.