
Part 2: Analysis of U.S. Domestic Flight Delays



Student Number: 240680602

Module: ST2195 - Programming for Data Science

Component: Coursework Part 2

Academic Year: 2025 / 2026

Table of Contents

I. Introduction	5
II. Data Pipeline	5
III. Analysis A: Optimal Flight Times	5
III.1. Methodology	5
III.2. Key Findings	6
III.3. Robustness Check	6
IV. Analysis B: Aircraft Age and Delay Probability	6
IV.1. The Confounding Problem	6
IV.2. Model Design	7
IV.3. Results	7
V. Analysis C: Predicting Flight Diversions	8
V.1. The Rarity Problem and Modelling Choices	8
V.2. Feature Importance (2004)	8
V.3. Model Stability Across Years	9
VI. Conclusion	9
References	11

List of Figures

Figure 1 - Qa: Delay probability heatmap (Day $\times$ Time Block) and robustness bar chart, 2004.....	6
Figure 2 - Qb: Delay probability by exact aircraft age (scatter, left) and by 5-year cohort (bar chart, right), 2004.	7
Figure 3 - Qc: Logistic regression coefficients for $P(\text{Diverted} = 1)$, 2004.....	8
Figure 4 - Qc: Diversion model coefficients across years (2003-2007)	9

List of Tables

Table 1 - Adjusted odds ratios and model performance for Question B, by year.	7
Table 2 - Diversion model performance vs naive baseline, by year.	9

I. Introduction

The ASA Data Expo dataset contains all U.S. domestic scheduled flights from 2003 to 2007. This dataset comprises around 35 million flights [1] with departure and arrival time, carrier codes, tail numbers, delay figures, as well as diversion flags. This report aims to answer three questions based on this dataset:

- **Optimal Timing:** What day of the week and departure time block result in the lowest arrival delay probability?
- **Aircraft Age:** Does flying on an older aircraft increase the arrival delay risk significantly?
- **Diversion Prediction:** Is it possible to predict flights with a higher risk of diversion using a statistical model? What factors are significant for this prediction?

The analysis for this problem was performed using both Python (Jupyter Notebook) and R (R Markdown). The steps taken for analysis are identical for both environments to ensure that the findings are not environment-dependent. All figures in this report are based on the 2004 data year, as it represents all data from 2003 to 2007.

II. Data Pipeline

The number of rows in the dataset was in the several millions per year, so efficiency was a concern from the beginning. Hence, only the columns needed were loaded, and each year was processed individually.

The cleaning of the data was focused on a few practical problems. Flights where arrival delay was a requirement but was missing (usually due to cancellation or diversion) were filtered out. Diverted flights were also set aside for special analysis. Clearly invalid air times, i.e., those which were zero or negative or exceeded total elapsed flight time, were changed to missing instead of throwing out the flight. Tail numbers which were not standard, e.g., “UNKNOWN,” were also changed to missing to prevent wrong matches in aircraft data.

A few new features were derived. Aircraft age was derived by calculating from manufacture year. Unrealistic values were discarded. Departure times were grouped into six-time blocks, from Night to Evening.

III. Analysis A: Optimal Flight Times

III.1. Methodology

For each year, the probability of an arrival delay, denoted as $P(\text{ArrDelay} > 0)$ was computed across all 42 day-of-week $\times$ time-block combinations. Cells with fewer than 30 observations were excluded before

identifying the optimal pairing. The analysis was repeated independently for each of the five years to test consistency.

III.2. Key Findings

The 2004 heatmap in **Figure 1** is a good example, with the darkest greens (lowest delay rate) dominating the early morning column from 05:00 to 08:00, whereas the afternoon peak from 16:00 to 20:00 and evening slots are deep red for all days. Sunday's early morning holds the lowest single-cell value at 27.7%, whereas Friday's afternoon peak holds a maximum value of 54.5%.

Saturday has the lowest delay rates across all individual days, and the early morning block is the top performer for each time in the five years combined. Friday afternoon is likely to have the highest delay rate for several reasons, including business travelers, connecting flights, and a full operating week's disruptions all happening at once.

III.3. Robustness Check

The right panel of **Figure 1** repeats the analysis with a more stringent definition of delay, where waiting times are defined as greater than 15 minutes (green bars). This definition arguably makes more sense, as such delays are likely to result in missing connections. Under both definitions, early morning remains the safest time block, with a stronger advantage using the more stringent definition. Not only do these flights have fewer delays, they also have shorter ones when they do occur. The optimal day-time combination remains the same for both definitions in 2004 as well as in most of the five years.

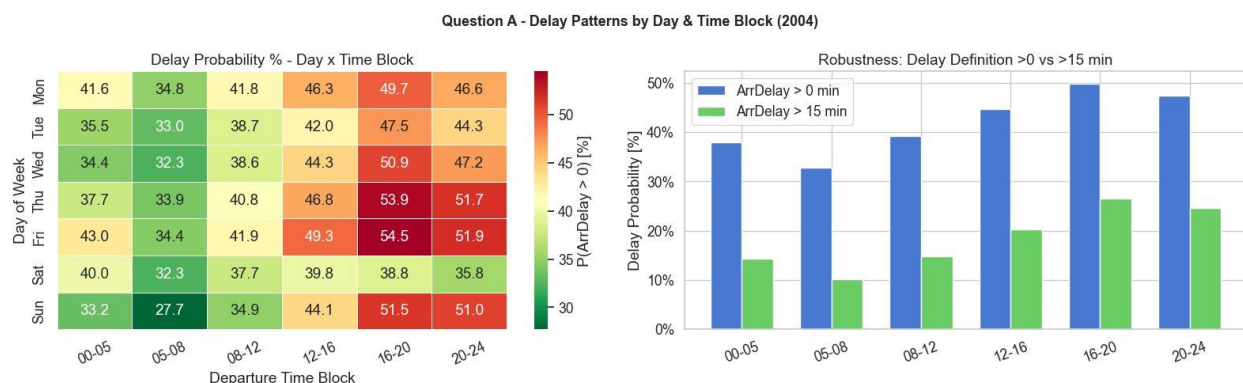


Figure 1 - Qa: Delay probability heatmap (Day $\times$ Time Block) and robustness bar chart, 2004.

IV. Analysis B: Aircraft Age and Delay Probability

IV.1. The Confounding Problem

It is not a random distribution of old planes on routes; low-fare airlines use old planes for smaller routes, and major airlines use newer planes for premium routes. A comparison of delay rates between age groups would be a straightforward method. In order to analyze the effect of age, a logistic regression [3] model

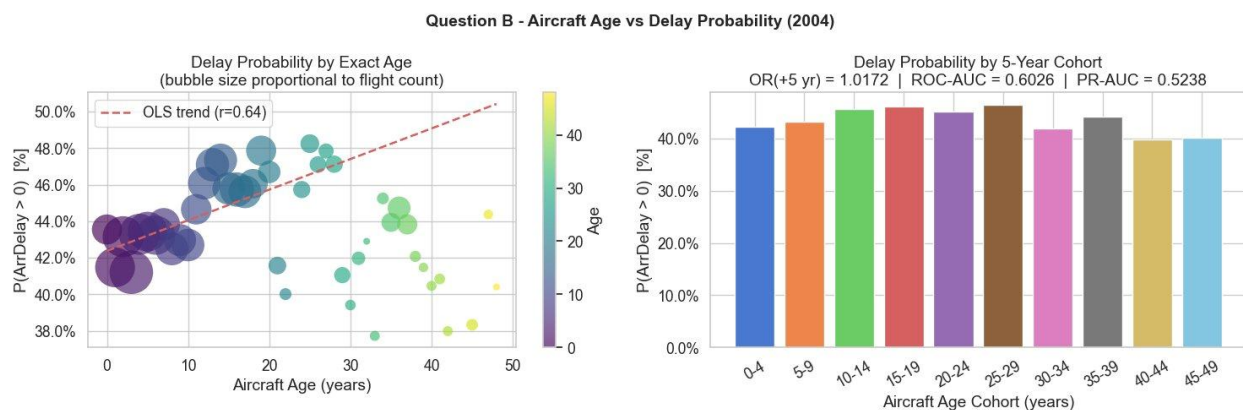
was used. In addition to age, flight month, day of week, departure time, distance, airline carrier, and origin/destination coordinates were used as control variables.

IV.2. Model Design

The model predicts a flight delay given several variables. A stratified subsample of up to 300,000 rows per year was used for computation. An 80/20 train-test split of data was used, and performance was measured using **ROC-AUC** and **PR-AUC** on 20% of the data.

IV.3. Results

The left-hand panel of **Figure 2** displays the raw relationship between aircraft age and probability of delay. Bubble sizes are standardized according to the count of flights. The OLS regression line displays a positive relationship between aircraft age and probability of delay. The regression line increases with aircraft age. In the year 2004, the r-value is equal to 0.64. The right-hand panel displays the five-year cohorts with the adjusted regression results superimposed.



As shown in **Table 1**, the adjusted odd ratios for a five-year increase in aircraft age increase from 1.01 in the year 2003 to 1.04 in the year 2007. In essence, the **odds** of experiencing a delay increase by 1% to 4% when flying on a five-year older aircraft, holding all other scheduling and routing factors constant. The ROC AUC is between 0.59 and 0.63, indicating that aircraft age is only a small contributor to the probability of delay. Scheduling effects play a larger part.

Table 1 - Adjusted odds ratios and model performance for Question B, by year.

Year	OR per +5 Years Age	ROC-AUC (Test)	PR-AUC (Test)
2003	1.01	0.59	0.47
2004	1.02	0.60	0.52
2005	1.02	0.61	0.54
2006	1.03	0.62	0.55
2007	1.04	0.63	0.58

V. Analysis C: Predicting Flight Diversions

V.1. The Rarity Problem and Modelling Choices

Diversions are rare, with an increase from 0.18% in 2003 to 0.24% in 2007. The model would have an accuracy of over 99.8% for predicting “not diverted” for each flight. Two strategies were used to counter this: **class weighting** is used to penalize missed diversions during training, and **PR-AUC** is used as the primary metric since it is more sensitive to rare event performance than **ROC-AUC** [4].

It should be noted that while class-weighting improves the model's ability to identify rare diversions (reflected in the improved PR-AUC), it naturally distorts the model's calibration. Consequently, the output probabilities serve as relative risk rankings rather than perfectly calibrated, real-world probabilities [4].

V.2. Feature Importance (2004)

In **Figure 3**, the top 20 coefficients for the year 2004 are depicted. The major factors influencing the model were the carriers: Alaska Airlines (AS), ExpressJet (XE), and SkyWest (OO) have the largest coefficients. Flight distance is another important factor. As would be expected, longer distances were associated with a greater risk of diversion.

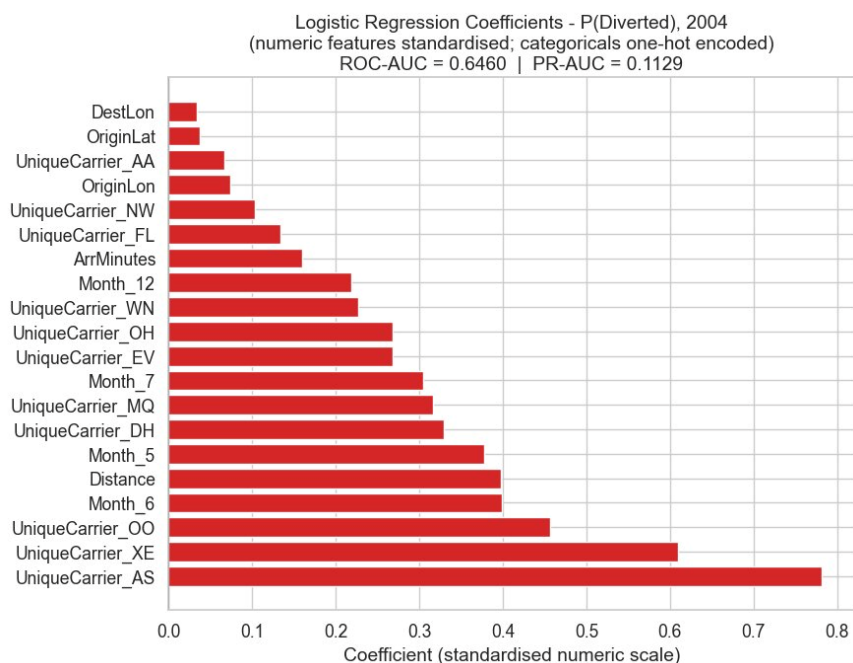


Figure 3 - Qc: Logistic regression coefficients for $P(\text{Diverted} = 1)$, 2004.

The strong coefficients for Alaska Airlines (AS), ExpressJet (XE), and SkyWest (OO) likely reflect their operational realities. Alaska Airlines frequently operates in harsh, unpredictable weather conditions, while ExpressJet and SkyWest operate smaller regional jets which are more vulnerable to weather disruptions and air traffic control constraints than large commercial airliners [2].

V.3. Model Stability Across Years

The multi-year coefficient plot in **Figure 4** plots numeric features from 2003 to 2007. Distance (brown) again tops the list as the most important feature, which remains stable around 0.40 except for a spike in 2006. Scheduled arrival time (blue) steadily increases, implying that flights scheduled to arrive later in the day are increasingly associated with diversion risk. Departure time (green) consistently has a negative coefficient, implying that early departures are associated with less diversion risk.

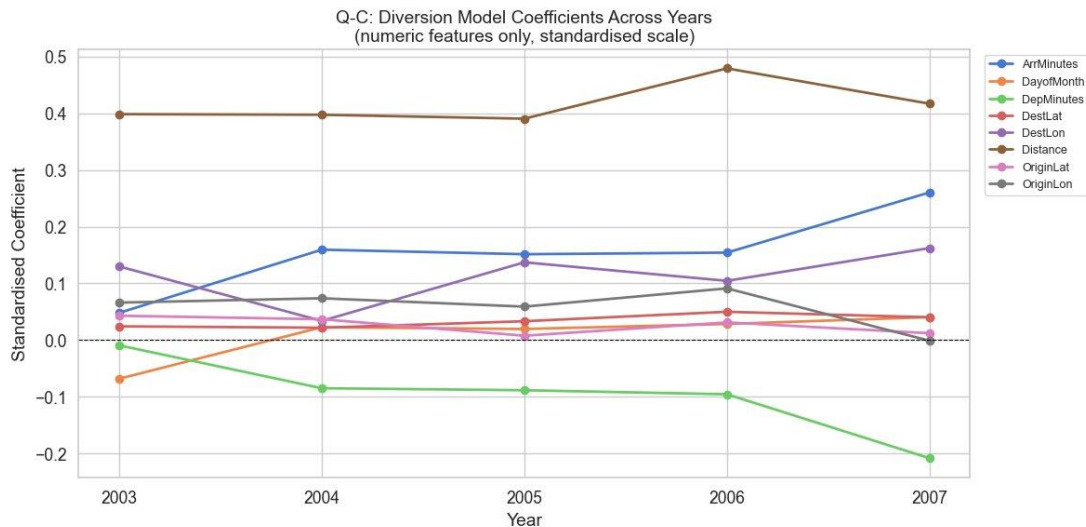


Figure 4 - Qc: Diversion model coefficients across years (2003-2007)

As can be seen from the results summarized in **Table 2**, the proposed model consistently outperforms the naive approach. In fact, in the year 2007, the performance of the proposed approach with a **PR-AUC** of 0.159 is nearly double that of the naive approach with a **PR-AUC** of 0.086.

Table 2 - Diversion model performance vs naive baseline, by year.

Year	Diversion Rate	Baseline AUC	PR-Model PR-AUC	ROC-AUC
2003	0.18%	0.057	0.097	~0.65
2004	0.19%	0.060	0.113	~0.65
2005	0.20%	0.065	0.125	~0.66
2006	0.22%	0.075	0.142	~0.67
2007	0.24%	0.086	0.159	~0.68

VI. Conclusion

Based on the analysis of domestic aviation in the USA during 2003 to 2007, a number of key factors associated with delays and diversions have been identified, which were analyzed using a substantial data set in each of the areas of focus. As a result of considering the findings on flight time, aircraft age, and forecasting techniques, several patterns of airline operations have become noticeable.

Firstly, flight time is essential for minimizing delays as they depend considerably on when a certain aircraft departs from the airport. Early-morning departures, in particular, tend to take place in conditions where flights will face fewer delays, whereas the likelihood of delays increases significantly towards the afternoon and evening, particularly during Friday's peak hours.

Secondly, it turned out that the age of aircraft slightly increases the risk of delay. However, the influence exerted by the age of an airplane on this factor is much lower compared to other aspects of operation. Using logistic regression, it became clear that the most influential factor is the identity of the airline rather than the age of an aircraft.

Finally, regarding diversions, which are a relatively rare phenomenon, the probability of such an incident could hardly be predicted. PR-AUC has proved to be a valuable metric because it performs better than the accuracy rate. Among the two factors found to influence predictions of such diversions, the distance and carrier identity should be highlighted.

To sum up, factors influencing airline delays and deviations can be largely attributed to operation-related aspects like departure time and the airline itself, and the plane's age is a less influential element. These findings provide a data-driven basis for understanding reliability in the domestic aviation network.

References

- [1] American Statistical Association (2009). *2009 ASA Data Expo*. <http://stat-computing.org/dataexpo/2009/>
- [2] Bureau of Transportation Statistics (2007). *Airline on-time statistics and delay causes*. U.S. Department of Transportation.
- [3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer.
- [4] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432.
- [5] Wickham, H., & Grolemund, G. (2017). *R for Data Science*. O'Reilly Media.