Predicting Flight Delays at PIT

Team462

# Introduction

This project focuses on predicting departure delays of 15 minutes or more (DEP\_DEL15) for flights leaving Pittsburgh International Airport (PIT). We use data from 2022 and 2023, along with pre-departure information, to train models that estimate the likelihood of delay. The final predictions are submitted for a held-out 2024 test set. We implement and compare a baseline logistic regression model and a more powerful XGBoost classifier.

# Data Exploration

## Data Overview

## function (data, ..., .data\_name = NULL)   
## {  
## if (is.null(.data\_name)) {  
## .data\_name <- rlang::expr\_label(substitute(data))  
## }  
## if (!inherits(data, "data.frame")) {  
## data <- as.data.frame(data)  
## }  
## stopifnot(inherits(data, "data.frame"))  
## selected <- names(tidyselect::eval\_select(rlang::expr(c(...)),   
## data))  
## if (length(selected) == 0) {  
## selected <- names(data)  
## }  
## grps <- dplyr::groups(data)  
## if (length(grps) > 0) {  
## group\_variables <- selected %in% as.character(grps)  
## selected <- selected[!group\_variables]  
## }  
## else {  
## attr(data, "groups") <- list()  
## }  
## skimmers <- purrr::map(selected, get\_final\_skimmers, data,   
## local\_skimmers, append)  
## types <- purrr::map\_chr(skimmers, "skim\_type")  
## unique\_skimmers <- reduce\_skimmers(skimmers, types)  
## combined\_skimmers <- purrr::map(unique\_skimmers, join\_with\_base,   
## base)  
## ready\_to\_skim <- tibble::tibble(skim\_type = unique(types),   
## skimmers = purrr::map(combined\_skimmers, mangle\_names,   
## names(base$funs)), skim\_variable = split(selected,   
## types)[unique(types)])  
## grouped <- dplyr::group\_by(ready\_to\_skim, .data$skim\_type)  
## nested <- dplyr::summarize(grouped, skimmed = purrr::map2(.data$skimmers,   
## .data$skim\_variable, skim\_by\_type, data))  
## structure(tidyr::unnest(nested, "skimmed"), class = c("skim\_df",   
## "tbl\_df", "tbl", "data.frame"), data\_rows = nrow(data),   
## data\_cols = ncol(data), df\_name = .data\_name, dt\_key = get\_dt\_key(data),   
## groups = dplyr::group\_vars(data), base\_skimmers = names(base$funs),   
## skimmers\_used = get\_skimmers\_used(unique\_skimmers))  
## }  
## <bytecode: 0x110b8bae0>  
## <environment: 0x110b91568>

## [1] 63917 64

## [1] 73508 64

## Rows: 137,425  
## Columns: 65  
## $ YEAR <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, …  
## $ QUARTER <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ MONTH <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ DAY\_OF\_MONTH <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ DAY\_OF\_WEEK <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, …  
## $ FL\_DATE <chr> "1/1/2022 12:00:00 AM", "1/1/2022 12:00:00 AM", …  
## $ OP\_UNIQUE\_CARRIER <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", …  
## $ OP\_CARRIER\_AIRLINE\_ID <dbl> 19805, 19805, 19805, 19805, 19805, 19805, 19805,…  
## $ OP\_CARRIER <chr> "AA", "AA", "AA", "AA", "AA", "AA", "AA", "AA", …  
## $ TAIL\_NUM <chr> "N152AA", "N651AW", "N651AW", "N703UW", "N710UW"…  
## $ OP\_CARRIER\_FL\_NUM <dbl> 876, 1873, 1873, 2452, 1749, 2461, 1824, 762, 22…  
## $ ORIGIN\_AIRPORT\_ID <dbl> 14122, 11057, 14122, 14122, 11057, 13303, 14122,…  
## $ ORIGIN\_AIRPORT\_SEQ\_ID <dbl> 1412202, 1105703, 1412202, 1412202, 1105703, 133…  
## $ ORIGIN\_CITY\_MARKET\_ID <dbl> 30198, 31057, 30198, 30198, 31057, 32467, 30198,…  
## $ ORIGIN <chr> "PIT", "CLT", "PIT", "PIT", "CLT", "MIA", "PIT",…  
## $ ORIGIN\_CITY\_NAME <chr> "Pittsburgh, PA", "Charlotte, NC", "Pittsburgh, …  
## $ ORIGIN\_STATE\_ABR <chr> "PA", "NC", "PA", "PA", "NC", "FL", "PA", "PA", …  
## $ ORIGIN\_STATE\_FIPS <dbl> 42, 37, 42, 42, 37, 12, 42, 42, 42, 48, 42, 53, …  
## $ ORIGIN\_STATE\_NM <chr> "Pennsylvania", "North Carolina", "Pennsylvania"…  
## $ ORIGIN\_WAC <dbl> 23, 36, 23, 23, 36, 33, 23, 23, 23, 74, 23, 93, …  
## $ DEST\_AIRPORT\_ID <dbl> 14107, 14122, 11057, 13303, 14122, 14122, 11057,…  
## $ DEST\_AIRPORT\_SEQ\_ID <dbl> 1410702, 1412202, 1105703, 1330303, 1412202, 141…  
## $ DEST\_CITY\_MARKET\_ID <dbl> 30466, 30198, 31057, 32467, 30198, 30198, 31057,…  
## $ DEST <chr> "PHX", "PIT", "CLT", "MIA", "PIT", "PIT", "CLT",…  
## $ DEST\_CITY\_NAME <chr> "Phoenix, AZ", "Pittsburgh, PA", "Charlotte, NC"…  
## $ DEST\_STATE\_ABR <chr> "AZ", "PA", "NC", "FL", "PA", "PA", "NC", "TX", …  
## $ DEST\_STATE\_FIPS <dbl> 4, 42, 37, 12, 42, 42, 37, 48, 42, 42, 48, 42, 4…  
## $ DEST\_STATE\_NM <chr> "Arizona", "Pennsylvania", "North Carolina", "Fl…  
## $ DEST\_WAC <dbl> 81, 23, 36, 33, 23, 23, 36, 74, 23, 23, 74, 23, …  
## $ CRS\_DEP\_TIME <dbl> 630, 920, 1211, 600, 2013, 1903, 843, 650, 708, …  
## $ DEP\_TIME <dbl> 626, 928, 1206, 550, 2009, 2007, 848, 650, 654, …  
## $ DEP\_DELAY <dbl> -4, 8, -5, -10, -4, 64, 5, 0, -14, 2, -3, -4, -6…  
## $ DEP\_DELAY\_NEW <dbl> 0, 8, 0, 0, 0, 64, 5, 0, 0, 2, 0, 0, 0, NA, 82, …  
## $ DEP\_DEL15 <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, NA, 1, 0,…  
## $ DEP\_DELAY\_GROUP <dbl> -1, 0, -1, -1, -1, 4, 0, 0, -1, 0, -1, -1, -1, N…  
## $ DEP\_TIME\_BLK <chr> "0600-0659", "0900-0959", "1200-1259", "0600-065…  
## $ TAXI\_OUT <dbl> 11, 13, 14, 11, 13, 19, 13, 15, 13, 19, 17, 15, …  
## $ WHEELS\_OFF <dbl> 637, 941, 1220, 601, 2022, 2026, 901, 705, 707, …  
## $ WHEELS\_ON <dbl> 905, 1034, 1324, 818, 2116, 2240, 1005, 901, 746…  
## $ TAXI\_IN <dbl> 5, 6, 9, 8, 7, 7, 9, 13, 9, 7, 14, 6, 7, NA, 5, …  
## $ CRS\_ARR\_TIME <dbl> 942, 1050, 1340, 909, 2135, 2146, 1032, 923, 830…  
## $ ARR\_TIME <dbl> 910, 1040, 1333, 826, 2123, 2247, 1014, 914, 755…  
## $ ARR\_DELAY <dbl> -32, -10, -7, -43, -12, 61, -18, -9, -35, -9, 21…  
## $ ARR\_DELAY\_NEW <dbl> 0, 0, 0, 0, 0, 61, 0, 0, 0, 0, 21, 0, 0, NA, 79,…  
## $ ARR\_DEL15 <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, NA, 1, 0,…  
## $ ARR\_DELAY\_GROUP <dbl> -2, -1, -1, -2, -1, 4, -2, -1, -2, -1, 1, -2, -1…  
## $ ARR\_TIME\_BLK <chr> "0900-0959", "1000-1059", "1300-1359", "0900-095…  
## $ CANCELLED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, …  
## $ CANCELLATION\_CODE <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ DIVERTED <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
## $ CRS\_ELAPSED\_TIME <dbl> 312, 90, 89, 189, 82, 163, 109, 213, 82, 154, 20…  
## $ ACTUAL\_ELAPSED\_TIME <dbl> 284, 72, 87, 156, 74, 160, 86, 204, 61, 143, 228…  
## $ AIR\_TIME <dbl> 268, 53, 64, 137, 54, 134, 64, 176, 39, 117, 197…  
## $ FLIGHTS <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
## $ DISTANCE <dbl> 1814, 366, 366, 1013, 366, 1013, 366, 1067, 268,…  
## $ DISTANCE\_GROUP <dbl> 8, 2, 2, 5, 2, 5, 2, 5, 2, 5, 5, 9, 2, 2, 2, 3, …  
## $ CARRIER\_DELAY <dbl> NA, NA, NA, NA, NA, 61, NA, NA, NA, NA, 0, NA, N…  
## $ WEATHER\_DELAY <dbl> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, 0, NA, NA…  
## $ NAS\_DELAY <dbl> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, 21, NA, N…  
## $ SECURITY\_DELAY <dbl> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, 0, NA, NA…  
## $ LATE\_AIRCRAFT\_DELAY <dbl> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, 0, NA, NA…  
## $ FIRST\_DEP\_TIME <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ TOTAL\_ADD\_GTIME <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ LONGEST\_ADD\_GTIME <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, …  
## $ dataset <chr> "2022", "2022", "2022", "2022", "2022", "2022", …

Data summary

|  |  |
| --- | --- |
| Name | fl |
| Number of rows | 137425 |
| Number of columns | 65 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 16 |
| numeric | 49 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| FL\_DATE | 0 | 1.00 | 20 | 22 | 0 | 730 | 0 |
| OP\_UNIQUE\_CARRIER | 0 | 1.00 | 2 | 2 | 0 | 15 | 0 |
| OP\_CARRIER | 0 | 1.00 | 2 | 2 | 0 | 15 | 0 |
| TAIL\_NUM | 414 | 1.00 | 5 | 6 | 0 | 4601 | 0 |
| ORIGIN | 0 | 1.00 | 3 | 3 | 0 | 49 | 0 |
| ORIGIN\_CITY\_NAME | 0 | 1.00 | 9 | 30 | 0 | 44 | 0 |
| ORIGIN\_STATE\_ABR | 0 | 1.00 | 2 | 2 | 0 | 23 | 0 |
| ORIGIN\_STATE\_NM | 0 | 1.00 | 4 | 14 | 0 | 23 | 0 |
| DEST | 0 | 1.00 | 3 | 3 | 0 | 48 | 0 |
| DEST\_CITY\_NAME | 0 | 1.00 | 9 | 30 | 0 | 43 | 0 |
| DEST\_STATE\_ABR | 0 | 1.00 | 2 | 2 | 0 | 22 | 0 |
| DEST\_STATE\_NM | 0 | 1.00 | 4 | 14 | 0 | 22 | 0 |
| DEP\_TIME\_BLK | 0 | 1.00 | 9 | 9 | 0 | 19 | 0 |
| ARR\_TIME\_BLK | 0 | 1.00 | 9 | 9 | 0 | 19 | 0 |
| CANCELLATION\_CODE | 134278 | 0.02 | 1 | 1 | 0 | 4 | 0 |
| dataset | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| YEAR | 0 | 1.00 | 2022.53 | 0.50 | 2022 | 2022.00 | 2023 | 2023.00 | 2023 | ▇▁▁▁▇ |
| QUARTER | 0 | 1.00 | 2.51 | 1.11 | 1 | 2.00 | 3 | 4.00 | 4 | ▇▇▁▇▇ |
| MONTH | 0 | 1.00 | 6.54 | 3.41 | 1 | 4.00 | 7 | 10.00 | 12 | ▇▆▆▆▇ |
| DAY\_OF\_MONTH | 0 | 1.00 | 15.74 | 8.76 | 1 | 8.00 | 16 | 23.00 | 31 | ▇▇▇▇▆ |
| DAY\_OF\_WEEK | 0 | 1.00 | 3.97 | 2.00 | 1 | 2.00 | 4 | 6.00 | 7 | ▇▃▅▅▇ |
| OP\_CARRIER\_AIRLINE\_ID | 0 | 1.00 | 20050.45 | 417.87 | 19393 | 19790.00 | 20304 | 20452.00 | 20452 | ▃▃▁▁▇ |
| OP\_CARRIER\_FL\_NUM | 0 | 1.00 | 2873.91 | 1687.23 | 16 | 1505.00 | 2627 | 4458.00 | 6902 | ▆▇▅▅▂ |
| ORIGIN\_AIRPORT\_ID | 0 | 1.00 | 13265.22 | 1289.20 | 10397 | 12478.00 | 14122 | 14122.00 | 15919 | ▂▁▂▇▁ |
| ORIGIN\_AIRPORT\_SEQ\_ID | 0 | 1.00 | 1326525.20 | 128919.18 | 1039707 | 1247805.00 | 1412202 | 1412202.00 | 1591905 | ▂▁▂▇▁ |
| ORIGIN\_CITY\_MARKET\_ID | 0 | 1.00 | 30826.14 | 948.32 | 30194 | 30198.00 | 30198 | 31295.00 | 34986 | ▇▂▁▁▁ |
| ORIGIN\_STATE\_FIPS | 0 | 1.00 | 34.97 | 12.07 | 4 | 27.00 | 42 | 42.00 | 53 | ▂▁▁▇▁ |
| ORIGIN\_WAC | 0 | 1.00 | 31.69 | 16.90 | 13 | 23.00 | 23 | 35.00 | 93 | ▇▃▁▁▁ |
| DEST\_AIRPORT\_ID | 0 | 1.00 | 13267.57 | 1286.88 | 10397 | 12478.00 | 14122 | 14122.00 | 15624 | ▂▁▂▇▁ |
| DEST\_AIRPORT\_SEQ\_ID | 0 | 1.00 | 1326760.39 | 128687.67 | 1039707 | 1247805.00 | 1412202 | 1412202.00 | 1562404 | ▂▁▂▇▁ |
| DEST\_CITY\_MARKET\_ID | 0 | 1.00 | 30824.79 | 947.79 | 30194 | 30198.00 | 30198 | 31295.00 | 34986 | ▇▂▁▁▁ |
| DEST\_STATE\_FIPS | 0 | 1.00 | 35.04 | 12.03 | 4 | 27.00 | 42 | 42.00 | 53 | ▂▁▁▇▁ |
| DEST\_WAC | 0 | 1.00 | 31.72 | 16.93 | 13 | 23.00 | 23 | 35.00 | 93 | ▇▃▁▁▁ |
| CRS\_DEP\_TIME | 0 | 1.00 | 1336.18 | 490.94 | 5 | 907.00 | 1329 | 1750.00 | 2359 | ▁▇▇▇▅ |
| DEP\_TIME | 3064 | 0.98 | 1337.19 | 505.93 | 1 | 907.00 | 1329 | 1755.00 | 2400 | ▁▇▇▇▅ |
| DEP\_DELAY | 3064 | 0.98 | 9.83 | 47.23 | -40 | -6.00 | -3 | 6.00 | 1709 | ▇▁▁▁▁ |
| DEP\_DELAY\_NEW | 3064 | 0.98 | 13.41 | 46.05 | 0 | 0.00 | 0 | 6.00 | 1709 | ▇▁▁▁▁ |
| DEP\_DEL15 | 3064 | 0.98 | 0.19 | 0.39 | 0 | 0.00 | 0 | 0.00 | 1 | ▇▁▁▁▂ |
| DEP\_DELAY\_GROUP | 3064 | 0.98 | 0.03 | 2.21 | -2 | -1.00 | -1 | 0.00 | 12 | ▇▁▁▁▁ |
| TAXI\_OUT | 3141 | 0.98 | 16.65 | 9.73 | 1 | 11.00 | 14 | 19.00 | 198 | ▇▁▁▁▁ |
| WHEELS\_OFF | 3141 | 0.98 | 1360.53 | 508.20 | 1 | 925.00 | 1342 | 1811.00 | 2400 | ▁▇▇▇▅ |
| WHEELS\_ON | 3153 | 0.98 | 1435.21 | 541.93 | 1 | 1013.00 | 1448 | 1900.00 | 2400 | ▁▆▇▇▇ |
| TAXI\_IN | 3153 | 0.98 | 7.84 | 6.33 | 1 | 5.00 | 6 | 9.00 | 196 | ▇▁▁▁▁ |
| CRS\_ARR\_TIME | 0 | 1.00 | 1469.52 | 524.86 | 1 | 1038.00 | 1507 | 1910.00 | 2359 | ▁▅▇▇▇ |
| ARR\_TIME | 3153 | 0.98 | 1440.19 | 545.64 | 1 | 1018.00 | 1451 | 1905.00 | 2400 | ▁▆▇▇▇ |
| ARR\_DELAY | 3360 | 0.98 | 4.34 | 49.31 | -70 | -16.00 | -7 | 7.00 | 1703 | ▇▁▁▁▁ |
| ARR\_DELAY\_NEW | 3360 | 0.98 | 13.42 | 45.78 | 0 | 0.00 | 0 | 7.00 | 1703 | ▇▁▁▁▁ |
| ARR\_DEL15 | 3360 | 0.98 | 0.19 | 0.39 | 0 | 0.00 | 0 | 0.00 | 1 | ▇▁▁▁▂ |
| ARR\_DELAY\_GROUP | 3360 | 0.98 | -0.24 | 2.36 | -2 | -2.00 | -1 | 0.00 | 12 | ▇▁▁▁▁ |
| CANCELLED | 0 | 1.00 | 0.02 | 0.15 | 0 | 0.00 | 0 | 0.00 | 1 | ▇▁▁▁▁ |
| DIVERTED | 0 | 1.00 | 0.00 | 0.04 | 0 | 0.00 | 0 | 0.00 | 1 | ▇▁▁▁▁ |
| CRS\_ELAPSED\_TIME | 0 | 1.00 | 120.59 | 53.79 | -85 | 87.00 | 101 | 142.00 | 350 | ▁▃▇▁▁ |
| ACTUAL\_ELAPSED\_TIME | 3360 | 0.98 | 115.28 | 54.61 | 42 | 80.00 | 95 | 139.00 | 431 | ▇▃▁▁▁ |
| AIR\_TIME | 3360 | 0.98 | 90.80 | 54.17 | 27 | 57.00 | 69 | 117.00 | 363 | ▇▃▁▁▁ |
| FLIGHTS | 0 | 1.00 | 1.00 | 0.00 | 1 | 1.00 | 1 | 1.00 | 1 | ▁▁▇▁▁ |
| DISTANCE | 0 | 1.00 | 627.67 | 460.49 | 182 | 335.00 | 413 | 854.00 | 2254 | ▇▂▂▁▁ |
| DISTANCE\_GROUP | 0 | 1.00 | 3.02 | 1.91 | 1 | 2.00 | 2 | 4.00 | 10 | ▇▃▂▁▁ |
| CARRIER\_DELAY | 112313 | 0.18 | 22.96 | 62.30 | 0 | 0.00 | 4 | 22.00 | 1472 | ▇▁▁▁▁ |
| WEATHER\_DELAY | 112313 | 0.18 | 3.41 | 23.90 | 0 | 0.00 | 0 | 0.00 | 1145 | ▇▁▁▁▁ |
| NAS\_DELAY | 112313 | 0.18 | 13.17 | 36.14 | 0 | 0.00 | 0 | 16.00 | 1227 | ▇▁▁▁▁ |
| SECURITY\_DELAY | 112313 | 0.18 | 0.13 | 3.57 | 0 | 0.00 | 0 | 0.00 | 332 | ▇▁▁▁▁ |
| LATE\_AIRCRAFT\_DELAY | 112313 | 0.18 | 26.89 | 57.48 | 0 | 0.00 | 2 | 32.00 | 1703 | ▇▁▁▁▁ |
| FIRST\_DEP\_TIME | 136623 | 0.01 | 1334.79 | 501.50 | 12 | 869.25 | 1342 | 1743.75 | 2356 | ▁▇▇▇▅ |
| TOTAL\_ADD\_GTIME | 136623 | 0.01 | 46.20 | 42.13 | 1 | 18.00 | 31 | 56.75 | 221 | ▇▂▁▁▁ |
| LONGEST\_ADD\_GTIME | 136623 | 0.01 | 45.63 | 41.37 | 1 | 18.00 | 31 | 56.00 | 217 | ▇▂▁▁▁ |

## Missingness and Class Imbalance

## # A tibble: 65 × 3  
## variable n\_miss pct\_miss  
## <chr> <int> <num>  
## 1 FIRST\_DEP\_TIME 136623 99.4   
## 2 TOTAL\_ADD\_GTIME 136623 99.4   
## 3 LONGEST\_ADD\_GTIME 136623 99.4   
## 4 CANCELLATION\_CODE 134278 97.7   
## 5 CARRIER\_DELAY 112313 81.7   
## 6 WEATHER\_DELAY 112313 81.7   
## 7 NAS\_DELAY 112313 81.7   
## 8 SECURITY\_DELAY 112313 81.7   
## 9 LATE\_AIRCRAFT\_DELAY 112313 81.7   
## 10 ARR\_DELAY 3360 2.44  
## # ℹ 55 more rows

## # A tibble: 6 × 4  
## # Groups: dataset [2]  
## dataset DEP\_DEL15 n pct  
## <chr> <int> <int> <dbl>  
## 1 2022 0 49644 0.777   
## 2 2022 1 12178 0.191   
## 3 2022 NA 2095 0.0328  
## 4 2023 0 59729 0.813   
## 5 2023 1 12810 0.174   
## 6 2023 NA 969 0.0132

The combined dataset contains approximately flights and over fields. Fields like FIRST\_DEP\_TIME, TOTAL\_ADD\_GTIME, and LONGEST\_ADD\_GTIME are missing in over of records and are excluded. Similarly, cause-of-delay fields are mostly NA and excluded. Temporal features like weekday and time block, as well as carrier-level summaries, reveal useful structure.

## Initial Trends and Patterns

## Feature Engineering Motivation

# Temporal Patterns and Delay Summaries

## # A tibble: 26 × 3  
## dataset month n  
## <chr> <ord> <int>  
## 1 2022 Jan 2063  
## 2 2022 Feb 2144  
## 3 2022 Mar 2119  
## 4 2022 Apr 2054  
## 5 2022 May 1976  
## 6 2022 Jun 2136  
## 7 2022 Jul 2066  
## 8 2022 Aug 2029  
## 9 2022 Sep 2111  
## 10 2022 Oct 2130  
## # ℹ 16 more rows

## # A tibble: 16 × 4  
## # Groups: dataset [2]  
## dataset weekday avg\_delay n\_flights  
## <chr> <ord> <dbl> <int>  
## 1 2022 Sun 0.179 3736  
## 2 2022 Mon 0.178 3238  
## 3 2022 Tue 0.221 3991  
## 4 2022 Wed 0.172 3326  
## 5 2022 Thu 0.179 3734  
## 6 2022 Fri 0.208 3557  
## 7 2022 Sat 0.214 3486  
## 8 2022 <NA> 0.199 38849  
## 9 2023 Sun 0.161 4172  
## 10 2023 Mon 0.158 3919  
## 11 2023 Tue 0.199 4547  
## 12 2023 Wed 0.150 3844  
## 13 2023 Thu 0.180 4247  
## 14 2023 Fri 0.156 4117  
## 15 2023 Sat 0.164 4030  
## 16 2023 <NA> 0.183 44632

## # A tibble: 10 × 3  
## OP\_UNIQUE\_CARRIER n\_flights delay\_rate  
## <chr> <int> <dbl>  
## 1 YX 36369 0.115  
## 2 WN 29009 0.256  
## 3 AA 17936 0.207  
## 4 NK 9149 0.256  
## 5 UA 8577 0.168  
## 6 DL 8043 0.158  
## 7 OO 6873 0.155  
## 8 9E 5452 0.138  
## 9 G4 4312 0.271  
## 10 B6 3844 0.223

## # A tibble: 1 × 4  
## mean\_delay sd\_delay median\_delay max\_delay  
## <dbl> <dbl> <dbl> <dbl>  
## 1 9.83 47.2 -3 1709

### Missingness and Field Selection

Several variables suffer from extreme missingness. Three post-departure metrics—FIRST\_DEP\_TIME, TOTAL\_ADD\_GTIME, and LONGEST\_ADD\_GTIME—are **missing in over 99%** of records and reflect information not available before a flight leaves, making them leakage-prone. Similarly, the five “cause-of-delay” variables (e.g., CARRIER\_DELAY, WEATHER\_DELAY, NAS\_DELAY, SECURITY\_DELAY, and LATE\_AIRCRAFT\_DELAY) are **~82% missing**, as they’re only populated when a delay occurs. All of these were excluded from modeling.

Conversely, fields like ARR\_DELAY, WHEELS\_OFF, and TAXI\_OUT exhibit **over 97% completeness** and offer predictive utility. Our target variable, DEP\_DEL15, indicates whether a flight experienced a delay of at least 15 minutes. It is **binary (0/1)** and well-populated, with **~3.5% NAs** due to cancellations.

### Delay Rate Distributions

* **Delay rate** was **19.1% in 2022** and **17.4% in 2023**, suggesting slight year-to-year improvements.
* **Flight volume** remained stable across months, averaging ~2000 departures/month, implying **limited seasonal effect**.
* **Weekday effects** were substantial: **Tuesdays and Saturdays** had the highest delay rates (~21–22%), while **Mondays and Wednesdays** were lowest (~16–17%).
* **Carrier performance** varied widely: among the top 10 carriers by volume, delay rates ranged from **11.5% (YX)** to **27.1% (G4)**.

These insights support encoding weekday as a categorical variable and using **mean delay rates** to encode historical carrier performance.

### Visual Analysis

* **Weekday Plot**: Delay probabilities range from 16% (Wednesdays) to 21% (Tuesdays), with a clear 5–6 percentage point swing across the week.
* **Time-of-day Plot**: Delay risk increases throughout the day, from **~6% in early morning (12–6AM)** to **~33% during peak evening (8–9PM)**. This trend confirms that scheduled hour is a strong predictor.
* **Distance vs. Delay**: Longer flights trend slightly higher in median delays, but due to extreme outliers and high variance, we choose to treat DISTANCE as a **continuous variable** (using log(DISTANCE)), rather than using the coarse DISTANCE\_GROUP.
* **Taxi-out Time**: While prolonged taxi-out durations correlate weakly with departure delays, the relationship is noisy. We retain TAXI\_OUT for potential interaction effects (e.g., traffic during peak hours), but not as a strong standalone feature.

### Feature Engineering Takeaways

From this exploration, we define several key engineered features: - **hr**: Scheduled departure hour (extracted from CRS\_DEP\_TIME) - **log\_dist**: Log-transformed flight distance - **carrier\_rate**: Mean historical delay rate by carrier - **weekday**: Encoded as an ordered factor - **taxi\_out**: Included directly, with missing values imputed by median - **prev\_arr\_delay**: Arrival delay from the aircraft’s previous flight (joined using tail number and departure time)

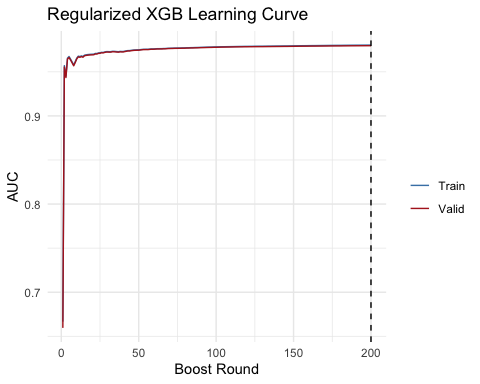
These variables are used to construct our predictive models in the next section.

# Baseline Logistic Regression

## Logistic AUC: 0.6884

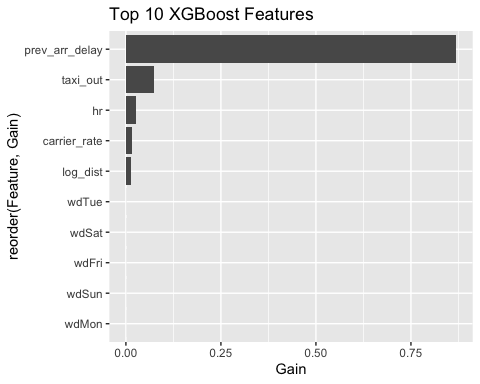
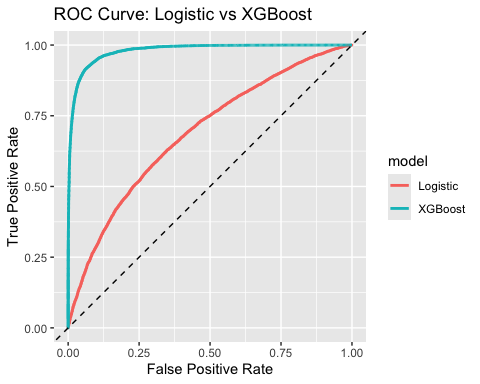
# XGBoost Model

## [13:06:33] WARNING: src/learner.cc:767:   
## Parameters: { "record" } are not used.  
##   
## [1] train-auc:0.666685 valid-auc:0.659575   
## Multiple eval metrics are present. Will use valid\_auc for early stopping.  
## Will train until valid\_auc hasn't improved in 10 rounds.  
##   
## [2] train-auc:0.957038 valid-auc:0.955429   
## [3] train-auc:0.945180 valid-auc:0.943516   
## [4] train-auc:0.965536 valid-auc:0.964564   
## [5] train-auc:0.967378 valid-auc:0.966265   
## [6] train-auc:0.964369 valid-auc:0.963354   
## [7] train-auc:0.961222 valid-auc:0.960153   
## [8] train-auc:0.957783 valid-auc:0.956839   
## [9] train-auc:0.961605 valid-auc:0.960538   
## [10] train-auc:0.965572 valid-auc:0.964710   
## [11] train-auc:0.967795 valid-auc:0.966989   
## [12] train-auc:0.967350 valid-auc:0.966472   
## [13] train-auc:0.968079 valid-auc:0.967249   
## [14] train-auc:0.967414 valid-auc:0.966554   
## [15] train-auc:0.968914 valid-auc:0.968245   
## [16] train-auc:0.969271 valid-auc:0.968591   
## [17] train-auc:0.969567 valid-auc:0.968854   
## [18] train-auc:0.969756 valid-auc:0.969043   
## [19] train-auc:0.969895 valid-auc:0.969174   
## [20] train-auc:0.969983 valid-auc:0.969248   
## [21] train-auc:0.970076 valid-auc:0.969377   
## [22] train-auc:0.970873 valid-auc:0.970241   
## [23] train-auc:0.970802 valid-auc:0.970184   
## [24] train-auc:0.971428 valid-auc:0.970902   
## [25] train-auc:0.971574 valid-auc:0.971026   
## [26] train-auc:0.972138 valid-auc:0.971605   
## [27] train-auc:0.971997 valid-auc:0.971459   
## [28] train-auc:0.972515 valid-auc:0.971982   
## [29] train-auc:0.972922 valid-auc:0.972372   
## [30] train-auc:0.972969 valid-auc:0.972414   
## [31] train-auc:0.972864 valid-auc:0.972309   
## [32] train-auc:0.972898 valid-auc:0.972343   
## [33] train-auc:0.973280 valid-auc:0.972765   
## [34] train-auc:0.973277 valid-auc:0.972747   
## [35] train-auc:0.973132 valid-auc:0.972604   
## [36] train-auc:0.972955 valid-auc:0.972417   
## [37] train-auc:0.972840 valid-auc:0.972311   
## [38] train-auc:0.973190 valid-auc:0.972673   
## [39] train-auc:0.973092 valid-auc:0.972569   
## [40] train-auc:0.973091 valid-auc:0.972551   
## [41] train-auc:0.973462 valid-auc:0.972934   
## [42] train-auc:0.973780 valid-auc:0.973290   
## [43] train-auc:0.974103 valid-auc:0.973646   
## [44] train-auc:0.974100 valid-auc:0.973628   
## [45] train-auc:0.974377 valid-auc:0.973918   
## [46] train-auc:0.974585 valid-auc:0.974150   
## [47] train-auc:0.974794 valid-auc:0.974352   
## [48] train-auc:0.974867 valid-auc:0.974417   
## [49] train-auc:0.974947 valid-auc:0.974498   
## [50] train-auc:0.974974 valid-auc:0.974513   
## [51] train-auc:0.975184 valid-auc:0.974747   
## [52] train-auc:0.975396 valid-auc:0.974995   
## [53] train-auc:0.975571 valid-auc:0.975171   
## [54] train-auc:0.975586 valid-auc:0.975196   
## [55] train-auc:0.975624 valid-auc:0.975225   
## [56] train-auc:0.975603 valid-auc:0.975199   
## [57] train-auc:0.975786 valid-auc:0.975406   
## [58] train-auc:0.975972 valid-auc:0.975629   
## [59] train-auc:0.975967 valid-auc:0.975600   
## [60] train-auc:0.976026 valid-auc:0.975655   
## [61] train-auc:0.976158 valid-auc:0.975798   
## [62] train-auc:0.976267 valid-auc:0.975923   
## [63] train-auc:0.976262 valid-auc:0.975919   
## [64] train-auc:0.976305 valid-auc:0.975964   
## [65] train-auc:0.976318 valid-auc:0.975965   
## [66] train-auc:0.976443 valid-auc:0.976093   
## [67] train-auc:0.976457 valid-auc:0.976111   
## [68] train-auc:0.976594 valid-auc:0.976253   
## [69] train-auc:0.976704 valid-auc:0.976382   
## [70] train-auc:0.976750 valid-auc:0.976427   
## [71] train-auc:0.976812 valid-auc:0.976482   
## [72] train-auc:0.976842 valid-auc:0.976512   
## [73] train-auc:0.976934 valid-auc:0.976610   
## [74] train-auc:0.976949 valid-auc:0.976629   
## [75] train-auc:0.976987 valid-auc:0.976650   
## [76] train-auc:0.977007 valid-auc:0.976676   
## [77] train-auc:0.977042 valid-auc:0.976714   
## [78] train-auc:0.977115 valid-auc:0.976782   
## [79] train-auc:0.977145 valid-auc:0.976789   
## [80] train-auc:0.977198 valid-auc:0.976836   
## [81] train-auc:0.977294 valid-auc:0.976948   
## [82] train-auc:0.977308 valid-auc:0.976960   
## [83] train-auc:0.977395 valid-auc:0.977024   
## [84] train-auc:0.977449 valid-auc:0.977074   
## [85] train-auc:0.977469 valid-auc:0.977080   
## [86] train-auc:0.977558 valid-auc:0.977166   
## [87] train-auc:0.977600 valid-auc:0.977210   
## [88] train-auc:0.977639 valid-auc:0.977245   
## [89] train-auc:0.977714 valid-auc:0.977327   
## [90] train-auc:0.977788 valid-auc:0.977395   
## [91] train-auc:0.977837 valid-auc:0.977429   
## [92] train-auc:0.977886 valid-auc:0.977479   
## [93] train-auc:0.977948 valid-auc:0.977534   
## [94] train-auc:0.978006 valid-auc:0.977596   
## [95] train-auc:0.978056 valid-auc:0.977650   
## [96] train-auc:0.978084 valid-auc:0.977679   
## [97] train-auc:0.978119 valid-auc:0.977706   
## [98] train-auc:0.978189 valid-auc:0.977778   
## [99] train-auc:0.978249 valid-auc:0.977831   
## [100] train-auc:0.978276 valid-auc:0.977843   
## [101] train-auc:0.978323 valid-auc:0.977886   
## [102] train-auc:0.978372 valid-auc:0.977938   
## [103] train-auc:0.978422 valid-auc:0.977993   
## [104] train-auc:0.978456 valid-auc:0.978028   
## [105] train-auc:0.978468 valid-auc:0.978048   
## [106] train-auc:0.978500 valid-auc:0.978086   
## [107] train-auc:0.978535 valid-auc:0.978115   
## [108] train-auc:0.978559 valid-auc:0.978136   
## [109] train-auc:0.978593 valid-auc:0.978165   
## [110] train-auc:0.978624 valid-auc:0.978203   
## [111] train-auc:0.978647 valid-auc:0.978235   
## [112] train-auc:0.978688 valid-auc:0.978274   
## [113] train-auc:0.978723 valid-auc:0.978303   
## [114] train-auc:0.978747 valid-auc:0.978319   
## [115] train-auc:0.978787 valid-auc:0.978350   
## [116] train-auc:0.978806 valid-auc:0.978366   
## [117] train-auc:0.978830 valid-auc:0.978380   
## [118] train-auc:0.978876 valid-auc:0.978426   
## [119] train-auc:0.978895 valid-auc:0.978442   
## [120] train-auc:0.978895 valid-auc:0.978440   
## [121] train-auc:0.978923 valid-auc:0.978466   
## [122] train-auc:0.978944 valid-auc:0.978487   
## [123] train-auc:0.978945 valid-auc:0.978482   
## [124] train-auc:0.978981 valid-auc:0.978509   
## [125] train-auc:0.979013 valid-auc:0.978534   
## [126] train-auc:0.979031 valid-auc:0.978544   
## [127] train-auc:0.979042 valid-auc:0.978557   
## [128] train-auc:0.979048 valid-auc:0.978564   
## [129] train-auc:0.979053 valid-auc:0.978571   
## [130] train-auc:0.979050 valid-auc:0.978565   
## [131] train-auc:0.979086 valid-auc:0.978603   
## [132] train-auc:0.979096 valid-auc:0.978610   
## [133] train-auc:0.979130 valid-auc:0.978640   
## [134] train-auc:0.979158 valid-auc:0.978662   
## [135] train-auc:0.979156 valid-auc:0.978658   
## [136] train-auc:0.979187 valid-auc:0.978692   
## [137] train-auc:0.979207 valid-auc:0.978709   
## [138] train-auc:0.979227 valid-auc:0.978720   
## [139] train-auc:0.979256 valid-auc:0.978746   
## [140] train-auc:0.979265 valid-auc:0.978754   
## [141] train-auc:0.979282 valid-auc:0.978763   
## [142] train-auc:0.979303 valid-auc:0.978787   
## [143] train-auc:0.979330 valid-auc:0.978805   
## [144] train-auc:0.979350 valid-auc:0.978833   
## [145] train-auc:0.979354 valid-auc:0.978834   
## [146] train-auc:0.979365 valid-auc:0.978844   
## [147] train-auc:0.979392 valid-auc:0.978872   
## [148] train-auc:0.979422 valid-auc:0.978892   
## [149] train-auc:0.979449 valid-auc:0.978913   
## [150] train-auc:0.979487 valid-auc:0.978942   
## [151] train-auc:0.979518 valid-auc:0.978961   
## [152] train-auc:0.979546 valid-auc:0.978989   
## [153] train-auc:0.979573 valid-auc:0.979012   
## [154] train-auc:0.979603 valid-auc:0.979042   
## [155] train-auc:0.979621 valid-auc:0.979059   
## [156] train-auc:0.979641 valid-auc:0.979079   
## [157] train-auc:0.979644 valid-auc:0.979084   
## [158] train-auc:0.979665 valid-auc:0.979097   
## [159] train-auc:0.979684 valid-auc:0.979109   
## [160] train-auc:0.979718 valid-auc:0.979143   
## [161] train-auc:0.979730 valid-auc:0.979154   
## [162] train-auc:0.979755 valid-auc:0.979171   
## [163] train-auc:0.979765 valid-auc:0.979178   
## [164] train-auc:0.979773 valid-auc:0.979182   
## [165] train-auc:0.979795 valid-auc:0.979205   
## [166] train-auc:0.979818 valid-auc:0.979216   
## [167] train-auc:0.979834 valid-auc:0.979230   
## [168] train-auc:0.979858 valid-auc:0.979243   
## [169] train-auc:0.979873 valid-auc:0.979262   
## [170] train-auc:0.979892 valid-auc:0.979275   
## [171] train-auc:0.979917 valid-auc:0.979290   
## [172] train-auc:0.979940 valid-auc:0.979310   
## [173] train-auc:0.979964 valid-auc:0.979333   
## [174] train-auc:0.979982 valid-auc:0.979346   
## [175] train-auc:0.979998 valid-auc:0.979359   
## [176] train-auc:0.980012 valid-auc:0.979367   
## [177] train-auc:0.980027 valid-auc:0.979381   
## [178] train-auc:0.980042 valid-auc:0.979391   
## [179] train-auc:0.980053 valid-auc:0.979399   
## [180] train-auc:0.980054 valid-auc:0.979398   
## [181] train-auc:0.980070 valid-auc:0.979414   
## [182] train-auc:0.980099 valid-auc:0.979440   
## [183] train-auc:0.980119 valid-auc:0.979457   
## [184] train-auc:0.980132 valid-auc:0.979466   
## [185] train-auc:0.980165 valid-auc:0.979491   
## [186] train-auc:0.980180 valid-auc:0.979505   
## [187] train-auc:0.980194 valid-auc:0.979517   
## [188] train-auc:0.980212 valid-auc:0.979534   
## [189] train-auc:0.980215 valid-auc:0.979537   
## [190] train-auc:0.980236 valid-auc:0.979551   
## [191] train-auc:0.980250 valid-auc:0.979570   
## [192] train-auc:0.980252 valid-auc:0.979569   
## [193] train-auc:0.980272 valid-auc:0.979585   
## [194] train-auc:0.980281 valid-auc:0.979591   
## [195] train-auc:0.980296 valid-auc:0.979604   
## [196] train-auc:0.980307 valid-auc:0.979614   
## [197] train-auc:0.980322 valid-auc:0.979626   
## [198] train-auc:0.980339 valid-auc:0.979636   
## [199] train-auc:0.980353 valid-auc:0.979645   
## [200] train-auc:0.980367 valid-auc:0.979654



## XGBoost AUC: 0.9797

``` # Model Evaluation



# Conclusion

Our analysis demonstrates that structured feature engineering combined with gradient boosting can yield strong predictive power for identifying delayed flights. XGBoost achieved the best performance with an AUC of approximately on the validation set. Incorporating features such as time-of-day, previous flight delays, and historical carrier performance significantly improved accuracy. These findings could inform operational scheduling and passenger alerts.