

# Final Project: Predicting US Flight Delays using Flight Characteristics and Weather Data

Oliver Ryder-Green

2022-11-26

# Introduction

To paraphrase a well-known idiom, ‘nothing is certain but death, taxes, and delayed flights.’ Flight delays are an inconvenience that almost all aviation passengers will experience at some point in their travels. Yet the burden of flight delays is not the same for all passengers. In particular, US passengers are not entitled to compensation for delays<sup>1</sup>. Yet, between 2013 and 2022, approximately one in every five flights from US airports was delayed by at least 15 minutes<sup>2</sup>. With more than 10 million scheduled passenger flights in the US each year<sup>3</sup>, the cost to passengers of flight delays is substantial. Indeed, the *Federal Aviation Administration* estimates that flight delays in the US from 2016 to 2019 cost passengers US\$62.6 billion in total. Short of relying on airlines to inform them of expected delays, there is little that US passengers can do to reliably avoid flight delays. Therefore, I apply the classification methods discussed in class to determine which factors best inform flight departure delays for domestic flights in the US. I also use regression methods discussed in class to consider the extent to which these factors are able to predict the length of departure delays that passengers experience.

Data from the *Bureau of Transportation Statistics* illustrates the prevalence of domestic flight delays. Among all US carriers, between 15–25% of departures were delayed from 2010 to 2022. Among the top US carriers<sup>4</sup>, the proportion of delayed flights is persistently higher than average. Evidently, some US carriers exhibit fewer than average flight delays (e.g., Delta Airlines), but top US carriers tend to demonstrate more frequent flight delays than the industry as a whole.

For US passengers, the fact that top US carriers experience more frequent departure delays may be of interest in trying to avoid delays. That said, more frequent delays at top airlines do not necessarily imply more severe (i.e., costly) delays for passengers. The *Bureau of Transportation Statistics* data shows that, among all US carriers, mean departure delay lengths were between 7 and 17 minutes on average from 2010 to 2022. Unfortunately, top US carriers again appear to perform worse than the industry as a whole. Without exception, top US carriers exhibit longer-than-average delays at some point in the period.

The *Bureau of Transportation Statistics* data also highlights that the frequency of delays varies by origin airport. In line with the above, around one in every five flights from a US airport is delayed. There are clearly some airports that persistently experience more frequent delays, over 50% of all flights in some cases, and some airports that experience few or no delays.

Several factors may determine whether or not a flight is delayed and for how long. The confluence of certain factors may also make delays more likely or lengthy. Moreover, some factors are hard to observe or nearly impossible to predict. The task of anticipating delays is, therefore, extremely difficult for passengers. That said, the data above suggests that some features that are readily observable for passengers may be useful in avoiding delays. If passengers face a choice of carriers and origin airports, they may be better able to avoid costly delays by choosing those that feature less frequent and shorter delays. The aim of this analysis is to identify factors that passengers might use to anticipate delays.

---

<sup>1</sup>source: [www.transportation.gov](http://www.transportation.gov)

<sup>2</sup>source: [www.bts.gov](http://www.bts.gov)

<sup>3</sup>source: [www.faa.gov](http://www.faa.gov)

<sup>4</sup>as measured by total number of flights serviced in 2010–2022.

## Data

To identify factors that inform whether a flight is delayed on departure and for how long, I use data from the *Bureau of Transportation Statistics' Airline On-Time Performance Data*<sup>5</sup> for January, March, September, and December in 2016, 2017, and 2018, respectively. The flight data contains 8777 observations on US domestic flights and 22 features, such as the flight date, origin airport, carrier, destination, distance, and other flight level characteristics. I combine this data with weather data from *Weather Underground*<sup>6</sup>. The weather data contains weather observations, such as average temperature, precipitation, and maximum wind speed, from corresponding airport weather stations on flight departure dates. The merged dataset contains 2687 observations and 28 features.

## Compiling and Cleaning

### Flight Data

I manually download *Bureau of Transportation Statistics' Airline On-Time Performance Data* for January, March, September, and December in 2016, 2017, and 2018, respectively. I import the data and compile using Pandas in Python (see corresponding Jupyter NB). The resulting dataset has 5,851,068 observations and 22 features. To make the dataset manageable, I draw a random subset (fraction=0.0015) from each month-year sample. The resulting dataset contains 8777 observations.

### Weather Data

I use web-scraping methods in Python (see corresponding Jupyter NB) to acquire historical weather data from *Weather Underground*. I use airport codes corresponding to origin airports for departures in the flight data to scrape historical weather data from airport weather stations. I acquire observations on temperatures, precipitation, sea level pressure, and max wind speed on the date of departure. The resulting dataset contains 6190 observations.

### Merged Data

I merge the flight and weather data on the date of departure and origin airport code. For the flight data, delays are identified as any flight departing more than 15 minutes late: `DepDel15=1` if delayed and `DepDel15=0` if otherwise. Delays measured in minutes are given by `DepDelay`. Since the supervised learning methods I utilise rely on the assumption that **target variables** do not have missing values, I drop observations if both `DepDel15` and `DepDelay` are missing because such observations contain no useful information for the analysis.

Since I am interested in predicting delays and delay lengths using flight characteristics and weather observations, I consider the proportion of missing values for these predictor variables. I find that there are no missing observations in the flight data. However, around 70% of observations have missing values for `Day Average Temp`, `High Temp`, `Low Temp`, `Max Wind Speed`, and `Sea Level Pressure` (see Jupyter NB). Moreover, around 90% of observations have missing values for `Precipitation`. Dropping observations missing weather data is costly in terms of observations. Yet, the weather data is of interest in prediction and is likely independent of many flight characteristics. I choose to omit observations that have missing values for `Day Average Temp`, `High Temp`, `Low Temp`, `Max Wind Speed`, and `Sea Level Pressure`. Further omitting observations that have missing values for precipitation is very costly.

---

<sup>5</sup>[www.transtats.bts.gov/](http://www.transtats.bts.gov/)

<sup>6</sup>[www.wunderground.com](http://www.wunderground.com)

## Feature Engineering

### Summary

