Aviation On-Time Performance Analysis

Joshua Poirier

Tuesday, January 12, 2016

Abstract

This study is inspired by the web application and study by Ritchie King and Nate Silver at fivethirtyeight.com titled *Which Flight Will Get You There Fastest?*. The web application can be found here. The accompanying documentation describes what they mean by "fastest" flights:

> Airline A says it will fly you from Seattle to Portland, Oregon, in 45 minutes, but actually takes 60 minutes. Airline B says it will fly the same route in 75 minutes, and actually takes 70 minutes. Which flight would you rather take? That seems easy. Airline A!

I extend this study by considering two cases:

- **Case 1:** Direct Flight - The consumer wants the fastest flight as defined in the aforementioned study
- **Case 2:** Connecting Flight - The consumer wants an on-time flight in order to avoid missing a tight connection

I hypothesize that the 3 best and worst airlines for the two cases are different. To show this, a multivariate regression model will be built for each case to account for confounding variables including **month**, **day of week**, **origin airport**, **destination airport**, and **time of day** in addition to the **airline**.

Introduction

This study utilizes aviation on-time performance data provided by the United States Department of Transportation here. The study analyzes **1812011** flights and includes hundreds of dummy variables in the culminating multivariate regression model comparing Cases 1 and 2. **Case 2** will use the *ARR_DELAY* field to compare against scheduled arrivals. **Case 1** will use the difference between the gate-to-gate flight time (*ACTUAL_ELAPSED_TIME* field) and the **target time**. I calculate target time using the simplified formula shown by Ritchie King and Nate Silver.

$$targettime = 0.117 * distance + 0.517 * (lonorigin - londest) + 43.2$$

This formula produces an estimated travel time in minutes. **distance** is the Great Circle Distance (shortest distance between two points on the surface of a sphere) and is provided in the data set. The coefficient **0.117** indicates that flights travel at 513 mph. **lonorigin** and **londest** represent the longitudes of the originating and destination airports. 30 seconds of flight time is added for every degree of westbound longitude travelled. The constant **43.2** indicates the time airlines budget for taxiing and inefficient routing (flying around severe weather).

In the interest of conciseness, the code used to produce these results is omitted from this report; however, it is available in the R Markdown file used to produce this report here. Let's take a look at the first few rows of the raw data!

MONTH	DAY_OF_WEEK	CARRIER	ORIGIN	DEST	CRS_DEP_TIME	ARR_DELAY
1	3	AA	JFK	LAX	0900	13
1	4	AA	JFK	LAX	0900	1
1	6	AA	JFK	LAX	0900	59

ACTUAL_ELAPSED_TIME	DISTANCE
384	2475
389	2475
379	2475

Feature Examination

In this section I take an independent look at the features expected to impact the on-time performance of aircraft.

Month

The *month* of the flight is expected to impact on-time performance due to the seasonality experienced by the United States. Airports experiencing harsh winter conditions are expected to perform more poorly during the winter months (December through March). To show this, let us establish a null hypothesis stating population means for each month are equal while the alternative hypothesis states the means are not equal (the indices 1 through 7 represent Monday through Sunday).

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$
$H_\alpha : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5 \neq \mu_6 \neq \mu_7$

For this, I compute the 95% confidence interval for the mean delays for each month (less the overall mean). If the confidence intervals do not include 0 we reject the null hypothesis.

Case 1:

Month	Lower	Mean	Upper
1	4.512049	4.654241	4.796433
2	1.411596	1.540049	1.668503
3	-2.264782	-2.165042	-2.065302
4	-3.466254	-3.365328	-3.264401

Case 2:

Month	Lower	Mean	Upper
1	0.8545432	0.9031737	0.9518043
2	0.9032465	0.9532444	1.0032422
3	-0.6113358	-0.5719228	-0.5325098
4	-1.0916424	-1.0531923	-1.0147423

Since the confidence intervals for each month do not all contain 0 we reject the null hypothesis and state that the mean delays for each month are not equal. This makes them strong candidates for features to be included in the regression model (further analysis will be performed during the modeling).

Day of the Week

Origin Airport

Destination Airport

Time of Departure

Airline