FlightBBQ model

November 30, 2015

1 Notations

All along, we will make use of the following notations:

- flighID is a unique identifier to each flight
- delay denotes by the time by which the flight was delayed. This variable may take negative values if the plane arrives before the announced arrival time.
- delayed denotes whether the flight was delayed
- departure denotes the departure airport
- departureState denotes the state of the departure airport
- arrival denotes the arrival airport
- arrivalState denotes the state of the arrival airport
- day denotes the day the flight was operated, a.k.a. Monday, Tuesday, and so on
- month denotes the month, a.k.a. Monday, Tuesday, and so on
- year denotes the year, a.k.a. January, February, etc
- season denotes the season, a.k.a. spring, summer, fall, winter
- carrier denotes the airline
- tailID denotes the tail number of the plane
- plane denotes the plane model, a.k.a. Airbus 320, Boeing 737, and so on
- manufacturer denotes the manufacturer of the plance, a.k.a. Airbus, Boeing, and so on
- distance denotes the distance for the flight
- timeAir denotes the time spent in the air
- cause denotes whether it was a weather delay, a security delay, and so on.

2 Model

From a mathematical perspective, all the variable we consider will be treated as random variables, apart from flighID the flight identifier. We will mainly work with conditional expectation. For example, the expected delay for a flight operated by carrier = United on a Thursday is

$$\mathbb{E}[\mathsf{delay} \mid \mathsf{carrier} = \mathsf{United}, \mathsf{day} = \mathsf{Thursday}]$$

Conditional expectations $\mathbb{E}[U|V]$ benefit from a range of interesting properties. The first one is that $\mathbb{E}[U|V]$ is the best approximation of U conditional to the knowledge of V in the following sense. If $\mathcal{F}$ is the set of all measurable functions $f \mid \mathbb{R} \to \mathbb{R}$:

$$\|U - \mathbb{E}[U|V]\|_{L_2} = \min_{f \in \mathcal{F}} \|U - f(V)\|_{L_2}$$

The equality means that among the quantities $f(V)$ we can build from the data V, $\mathbb{E}[U|V]$ is the closest one to U.

From a computational perspective, $\mathbb{E}[U|V = v]$ is extremely easy to compute if we have enough data. Just select entries where $V = v$, then average U over these entries. For example, if we want to compute the expected delay given a flight operated by United, we select all the flight corresponding to carrier = United and we compute the average of the column delay.

Warning This method of computation relies on the law of large numbers and therefore requires that we have enough data points when we compute the average. If we don't have enough data points or no point at all, we have to build a model that related delay to all other variables:

$$\mathsf{delay} = f(\mathsf{carrier}, \mathsf{day}, \ldots) + \text{noise}$$

f can implement a linear regression, a neural network, etc.