[DATA-05P] Flight delays in the Washington area

Miguel-Angel Canela, IESE Business School

June 20, 2016

Introduction

One of the points in the agenda of the Board of Directors of the Metropolitan Washington Airports Authority was the number of flight delays currently occurring in the area. 15 minutes delay was usually taken as a cause for a complain, and about 15% of the flights exceeded this threshold. Customers claims were piling up, with some opportunistic firms already advertising advice and legal support, and even promising rates of success.

The Chairman of the Board wanted to get advice on the extent to which flight delays could be predicted from simple, current data. With no time for an exhaustive study, they would use the data of January for the corridor Washington-New York City. They would include Baltimore airport to get a better picture, although it did not depend on the Washington Authority.

The data set

The data set covered 2,201 airplane flights carried out in January 2004, from the Washington DC area into the NYC area. The variables included are:

- The date in format “yyyy-mm-dd” (`date`).
- The scheduled time in format “HourMinute” (`schedtime`).
- The real departure time in format “HourMinute” (`deptime`).
- The airline carrier (`carrier`): Continental Airlines (CO), Atlantic Coast Airlines (DH), Delta Air (DL), Envoy Air (MQ), Comair (OH), ExpressJet Airlines (RU), United Air Lines (UA) or USAir (US).
- The departure airport (`origin`): Reagan, Dulles or Baltimore.
- The arrival airport (`dest`): Kennedy, Newark or LaGuardia.
- The distance travelled in miles (`distance`).
- The day of the week (`dayweek`).
- The day of the month (`daymonth`).
- The plane tail number (`tailnumber`)
- The weather conditions (`weather`): good or bad.

Source of the data: G Shmueli, NR Patel & PC Bruce (2010), *Data Mining for Business Intelligence*, Wiley.