Capstone project proposal

By Will Bishop

July 24, 2016

1. The problem

To predict the on-time performance of a flight based on:

- Origin and destination cities
- Airline
- Time of year, time of day, day of week
- Aircraft type and number of seats

These are the variables I have identified as likely to hold predictive power, although there are others in the data set.

The dependent variables are various measures of arrival delay, the most basic being simply the number of minutes between the scheduled and actual arrival times. There are also five variables that break down the number of minutes of delay into different causes: carrier delay, weather delay, National Air System delay, security delay, and late aircraft delay (meaning that the previous flight using the same aircraft arrived late). I expect to focus mainly on the overall delay time, but will experiment with predicting the individual causes of delay.

2. The client

The client is anyone wishing to book a flight who would like to estimate the likelihood of the flight arriving late. More specifically, when comparing two flight options based on price, a difference in expected delay time may offset some of the difference in price, depending on how important it is for the traveler to arrive on time. The idea is theoretically to develop an app that would allow travelers to compare flights based on on-time performance in addition to more traditional metrics such as price, comfort, and stopover time.

3. The data

The data come from the Bureau of Transportation Statistics' On-Time Performance table: http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time. They have data on every flight of major U.S. airlines in monthly tables. The latest available data is from May 2016 and the data seem to go back to at least 1990, although I have not yet verified that every single month between 1990 and 2016 is available and uncorrupted. The number of rows per monthly table is in the neighborhood of 440,000, meaning that 25 years of data would be over 100 million observations. This is a lot of airline data.

I will test all the analysis on a much smaller sample—probably one year, so that I can get a sense of seasonal effects—before running code on the large data set.

At this point I have no plans to incorporate weather data. The BTS data should capture all variance in past weather by including the exact locations and times of departure and arrival, and I think most flights are booked far enough in advance that future weather cannot be predicted beyond the general patterns for the region and season. However, there may be exceptions (sometimes a storm is known in advance, for instance) so it's possible weather data could be useful.

4. The approach

The first stage of the project will be exploratory data analysis to determine which variables appear most correlated with on-time performance. The next stage will be testing various models to see which have the most predictive power while trying to avoid overfitting. I envision trying linear regression and *k* nearest neighbors with different sets of variables. I am not sure whether this problem is suitable for other machine techniques such as naïve Bayes or decision trees, but will take advice on that.

5. The product

I usually explain best by writing, so I'll include a paper explaining what I did and the conclusions, and a set of graphs as an attachment. Of course, all the code will also be available.