Capstone project ideas

Will Bishop

July 5, 2016

The first two project ideas are based on my interest in the airline industry. I have found an enormous trove of airline industry data at sources such as:

- Bureau of Transportation Statistics: http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/subject_areas/airline_information/index.html
- FAA: https://www.faa.gov/data_research/
- Airline Data Project: http://web.mit.edu/airlinedata/www/default.html

The two ideas look at the industry from different (and possibly competing) perspectives: an airline trying to maximize its revenue, and customers trying to get the best experience.

Idea 1

The goal is to model the **load factor** of a flight: the number of passengers divided by the number of available seats. It would consider the origin and destination cities, price, carrier, time of year, and other available variables. As in many projects, the idea would be to test several different machine learning techniques and determine which one best predicts load factor without overfitting.

Right now, the best data set I have found for this idea is "T-100 Domestic Segment" from BTS, which (as I understand it) shows totals or averages of a given flight for a time period, such as a month or quarter, rather than every individual flight. I think this would make a very interesting project, but it's possible there's other data at an even more granular level—these airline data sets have a lot to explore.

The client would be the airline, which presumably (all else held equal) wants to avoid empty seats. It's *possible* the project could also get into airline revenue/profitability data to see how strongly that correlates with load factor, but I haven't investigated the data enough to know whether that will work.

Idea 2

The goal is to predict the **on-time performance** of a flight over a period of time. Again it would consider the origin and destination cities, carrier, time of year, and other available variables, and test different machine learning techniques to come up with the best model. The data would be the On-Time Statistics and Delay Causes from BTS. If I wanted to use price as a predictive variable, I would have to merge the on-time statistics with fare data, which may be feasible.

The most obvious client for this project is a customer, either an individual or a business trying to arrange corporate travel. The client could also be a travel agency (though these are fast disappearing), or an app

developer who wants to help customers predict which flights will have the best on-time performance for the price.

The third project idea is based on two of my other major interests: elections in my home state of Alaska (elections.alaska.gov) and Census demographic data (factfinder2.census.gov).

Idea 3

Alaska has 438 election precincts. The goal is to look at the data from those precincts in the 2014 Alaska elections, figure out which precincts are the most similar overall in their voting patterns, then match to Census block-level demographic data and model areas' similarity in voting based on their demographic characteristics. This would be at a far more granular level than most studies of nationwide voting behavior that I know about. The main challenge (aside from the actual modeling) would be matching Census blocks to precincts. Alaska's small population and my familiarity with the geography of many of its larger towns provide an advantage, but I still do not know how feasible this will be.

The client for this project could be a political strategist for an Alaskan campaign, looking for micro-level voter targeting strategies. The client could also be a pollster trying to figure out proper geographic weighting in a notoriously hard-to-poll state.