Trevor Hudson DAT_BOS_06

In the age of social media, nuanced issues quickly become binary as people reinforce their preexisting biases with whichever facts are convenient for this end. Not only are some facts ignored to avoid the cognitive dissonance of being faced with events that challenge an entrenched worldview, but it can be impossible to obtain facts or data that describe widespread issues as a whole.

I wanted to understand police shootings. How many of those killed are unarmed? How does race factor in? In which cases is a police officer more likely to face rebuke or punishment for killing an unarmed suspect? Seeing as the review process that leads to a suspension, firing, or acquittal is not transparent, I wanted to create a model for determining whether an officer will face retribution. Most of all, I wanted to learn from the data and identify some of the interesting trends that are indistinguishable until you look at a dataset as a whole.

I found my dataset on Crowdflower's website. The data was gathered by workers on Amazon Mechanical Turk who individually analyzed news stories. There are many "unclear" cells in each of the feature columns, where definitive data was not available. As this is more of an exploratory exercise, I chose to do logistic regression so that my output would be the probability that an officer will face suspension or be fired, given an input of an array of binary predictors.

Cleaning the data involved removing certain columns that could be deduced from other more categorically conducive features. I then had to divide the features into categorical and non-categorical, keeping them separate so that I could make binary dummy variables from the categorical, multiple choice derived variables previously in place. Before running getting the binary values, I did a bit of string manipulation, specifically getting rid of the year in the month columns so that I could use each month as a categorical feature in which, for example, March 2013 and March 2014 are considered in the same category.

I filled the blank cells with 'Unclear' and then ran getDummies. I then added the linear feature back into the data frame. For this feature, I filled blank values with the mean of the entire column.

At this point, I was ready to break up my dataframe into predictive features and the predicted result. Here I experimented with SelectKBest to try to reduce my features and avoid overfitting. In creating my model, a Linear Regression model proved the best performing, having tried a Decision Tree, Random Forest, and Extra Trees Classifier.

This was certainly an iterative process and the steps described were by no means sequential. Though I got a decent model from the algorithm I applied, I don't feel like I really learned or brought to light what I wanted to discover. Far too much of the data were null values or labeled "unclear". I still do feel that there are still

insights to be had, and I've been focusing on getting better at Pandas to bring these to my attention.

The model of police suspension/termination, coupled with an analysis of the features that went into it, could help people understand both the threat of harsh consequences faced by police when in pursuit of suspects, and the dangers that certain segments of the population endure at the hands of law enforcement. Hopefully, by exploring the data as a whole, we can have enough information to ignore our biases and see these problems for what they are, multifaceted and nuanced as they may be.