ECON 9000

Ryan Pregitzer

May 4, 2019

Take Home Exam

The Data

The data set we were tasked with studying for this exam came from the Chicago Police crime database (<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>), which includes crime data from 2001 to the present, minus the past 7 days from whenever I visited the site and downloaded the data (May 3, 2019). The table includes the Illinois crime code in addition to descriptions of the crime. Several location variables are included: latitude/longitude, community code, beat, ward, district, etc. There are 77 communities across Chicago, and beats are assigned to individual police cars. Districts, then, are made up of 3-5 police sectors, which are, in turn, made up of 3-5 beats. Furthermore, the ‘location’ variable gives context to each crime: ‘abandoned warehouse,’ ‘residential’, ‘school’, etc.

For my analysis, I limited and edited my data set to make it more manageable and useful for my models. For one, I restricted my variables to date, year, crime description, location, beat, and community. I parsed the date variable to get month and hour variables, and I encoded location and crime description to make them usable in my models. I also limited the entire data set to the 2017-2019 years. I assumed that these years are representative of the rest of the data set; I felt confident in doing so because exogenous variables did not seem to affect these years. For example, the economy has remained relatively stable for the past few years, the country has not changed its war status significantly—although such a variable may not be significant in a regression—and other similar citywide or national events/changes seemed insignificant. Contrast this to a 2001-2002 data set, during which the U.S. sent troops to the Middle East or a 2008-2009 partial data set, in which the economy suffered a severe recession.

Analysis

My initial model was a simple, supervised ML model that took 75% of the data set and attempted to create a model, using the remaining 25% of the data to test the created model. The predicted vs actual for this model was 0.0082, so I tried using KFold to test using different sets of 75-25 for the supervised model. I began by splitting the data set into 4 parts, but my predicted vs actuals were very similar to 0.0082 (0.007, 0.0084, etc.). I then split the data into 5 sets so that I would be used 80% of the data to predict the other 20%, a percentage that might be more successful. This, too, was relatively inaccurate, giving me predicted vs actuals in a similar range, with values of 0.0069, 0.0092, 0.0073, etc.

Finally, I thought I would attempt to model my data using random forests. It took some tinkering, but I could see a 0.25 for predicted vs actuals in my results. While not impressive, this relationship was a significant improvement over 0.008. Unfortunately, I did not achieve this relationship in my final edits of my program. Earlier on in the program, I believe I had reduced the number of estimators to 5 (as compared to 10 or 11) and I got 0.25. My current program, however, failed to replicate these results. My computer kills the process running random forests, so I get no results from my models, or I will get an ‘unable to allocate x-million bytes of memory’ for this process. I attempted to upload my data set to the Palmetto Cluster to run my analysis on that, but the upload froze at 52% and moved no further. The cluster may have a file size limit of 1 GB. Regardless, I was unable to get results from random forests. (Side note: decision tree modeling gave me similar results). My work ultimately failed to find a relationship and successfully model crime in Chicago based on the inputs I had and other limitations imposed on my work. More powerful computers and an expansion on my current models or variables would likely produce working models and useful results at a later time.