Calista MacHarrie

Intro to Data Science

25 March 2020

Project Proposal

There is reporting that shows that “in most major cities in America, shooters who kill someone with a gun have more than a 1 out of 2 chance that their crime will go unpunished. If the victim survives, the shooter’s chance of getting away with it increases to 2 out of 3. Every year, hundreds — even thousands — of nonfatal shootings never even get assigned to a detective.” I would like to look at how shootings are investigated across the United State and what kinds of variables (location, race of victim, demographics of the city) affect the outcomes. The outcome will be binary - whether the case involving a shooting open or closed. The data set that I’m using involves shootings in the cases of different crimes (i.e. aggravated assault, homicide, etc) and I will decide if I will only use homicide data or additional crime data after looking more in depth at the data sets by city.

The data that I would like to use will be downloaded from this website: <https://www.thetrace.org/violent-crime-data/>. The data was collected via public records requests sent to local police and sheriff’s departments across the country. Each data set by state is organized somewhat differently so I would look through all the data sets for the sets with the most variables that I could use and the most years in order to have enough data points, maybe a couple hundred if possible. I think I’d use maybe 5-10 cities. I would recode, join and organize the data, dropping unnecessary variable and renaming others and then organizing by city year. I would also want to be able to add a variable that changes the city into geospatial data it in order to visualize the data based on city. I would also look for data sets that give demographics for the selected city over the years available to try to include additional variables.

For the machine learning part, I will split the dataset and run it through the various toolkits to see which features matter the most and what comes out in each toolkit. I would maybe leave out the last year of data when I divide the data to see the time series change. Other than that I don’t really don’t enough about the machine learning part to say what else I will do.

A successful project for me (because this stuff takes me forever!) might just be some version of wrangling, joining, using machine learning on and visualizing the data in the data sets provided. I will also try to add additional demographics data or other relevant city data if I can find that information and if the first part of the project isn’t incredibly time consuming for me. I think the obvious outcome is that race of the victim and the location of the crime affect whether a case is open or closed and a version of a finished project for me might just look like running some models to show that is the case and visualizing the outcomes. I will look for additional variables that I can add from other datasets though to expand on that.