DS4420 – Machine Learning & Data Mining II

Final Project

Ryan Liang

3/22/21

Motivation:

Mass shootings in the U.S. are a serious problem, with 2019 showing record high numbers of both fatalities and shootings in just the first half of the year (Figure 1). As we enter further into 2021, the effort to reduce the number of mass shootings has been aided negligibly, and almost nothing is getting better. The goal of the project will be developing key insights about the victims and also the perpetrators of mass shootings, and analyzing the mental health, race, location, and motivation (and any other important information) about the killers and where the mass shootings take place. When tackling a project of this magnitude, it's important to keep in mind the sensitivity of it, and how important it is to not to draw generalizations and conclusions about race and/or gender, as that is what fuels stereotypes and dangerously leads to racial profiling. However, it is also important to study and analyze key patterns in order to get a full comprehensive picture and to understand the problem better, especially when the issue is one of this caliber, and especially when enough resources haven't been allocated to studying the problem in the first place.

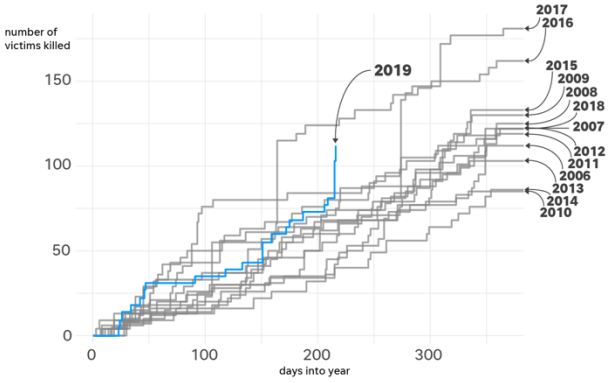

Figure 1

Data:

I will be using the US Mass Shootings dataset, which consists of mass shootings that occurred between the years of 1966 and 2017. It also includes information about the location of the shooting, the race of the killer, how many were injured, whether the killer was suffering from mental health issues, and other useful information.

Machine Learning Methods:

For the project, I will be most likely be using a Random Forest to compare some of the different machine learning models and understand how each one performs in relation to the data. In terms of the tasks to be completed, I will most likely be using classification models, but will try to see how I can use Regression to predict data in the future based on certain circumstances (predict the mass shootings in a certain city in a certain year, etc.). Another useful model I will be using is k-Means clustering model, to cluster mass shootings together based on several different features, such as number of fatalities, city, data, killer's race, etc. Grouping and classifying the data will be important for developing the key insights and underlying patterns of the data, which is why I will also be using a Naïve Bayes model as well. Any other useful classification models that I find along the way, I will implement for the data and compare performance.

Project:

The core work of the project will be finding the right machine learning models that will work best with the goal we are trying to reach. Before that, however, I will need to clean the data. There are some columns in the data that are not needed, as well as some that have too much information or are in formats that are not easy to work with (i.e., the date, the description). I will find a way to make this data easy to work with, and that will consist of good portion of the beginning of the project. Then, working with this newly cleaned dataset, I will implement different classification models and group the data based on different features and report the results and findings after each iteration, determining and analyzing patterns across iterations. It will all culminate in reporting and interpreting the most important findings and then developing actionable insights on how the findings are relevant to issue at hand and how the results can be used to help work towards solutions