a) identify a dataset you will use

The dataset we are going to use can be found at https://github.com/StanfordGeospatialCenter/MSA/blob/master/Data/Stanford_MSA_Database.csv.

This dataset contains data about mass shootings in America from 1966 to 2016. A mass shooting is defined as a shooting of 3 or more people killed in a public place from 1966-2012. After 2012, a mass shooting is defined as a shootingr in a public place with 4 or more victims killed. The data fields in this dataset are city/state, date, brief description, fatalities, injured, venue, mental health history, weapons obtained legally, where obtained, type of weapon/details, race, gender, and latitude and longitude of where the shooting occurred.

b) the outcome you are trying to predict, and what variables you will use to predict it;

The outcome we are trying to predict is the number of fatalities in total (civilian and law enforcement) and a possible motive based on the most significant factors. Based on the factors we deem to be relevant, we want to be able to predict who and how many people are at risk in a mass shooting.

c) motivation to your project -- as in the business or practical management use case of such a prediction;

The motivation for this project is that mass shootings have become a major problem in the United States. By analyzing a data set of this nature, it could be useful for schools and workplaces to track information on students or employees to calculate the risk of a suspicious person. Also by finding what factors are most impactful in mass shootings, these organizations could learn how the impact of mental health in particular plays into these tragedies and how important of a factor it is to address beforehand.

d) three methods you will use to analyze your question of interest;

One method we would use to analyze the question of interest is the lasso method. We will use that method because lasso is useful in variable selection and we want to see what factors make someone more at risk to commit a mass shooting.

Another method we will use will be multiple logistic regression. This is helpful because it multiple logistic regression is useful for classification problems. We want to classify the possible motive given certain factors. Along with the logistic method, we would interpret the results by looking at the odds ratio. This will be helpful in finding any relation in the factors in our data and how they impact the motive.

The third method we will use is a backwards stepwise selection. This again will be helpful because we want to find the variables that contribute to the likelihood someone would commit a mass shooting. Using this method, we can compare the variables that the backward stepwise algorithm decided to keep and the variables the lasso decided to choose. If they are the same, then we can have more confidence in our prediction, and if they are different, we can investigate why.

Lasso, elasticnet, randomforest

e) the names of the students who will be part of your group.

Teresa Gerhold

Merzia Culterywala

Josh Anderson

Nick O’Neill