Hello Dr. Linderman,

My name is Chris Goodhart and my partner’s name is Aryan Deorah. We are students at the Loudoun Academy of Science, where we have started a research project focused on determining a machine learning model that can predict rates of gun violence from gun laws. Our ultimate goal is to be able to make suggestions for policymakers as to which laws should be put in place to have the best impact on rates of gun violence. We are contacting you with the hope that you may be able and willing to assist us and offer insight as to how to better approach this problem.

So far, we have collected data for the gun laws present in a given state in a given year along with the age-adjusted rates of gun violence in that state and year. The database that we are using is attached. We have also collected data for median household income, poverty rates, and violent crime rates. We are working on obtaining more data for more controls such as race, unemployment, and population density. Regarding the socioeconomic factors (median household income, poverty rates, and unemployment), we understand that they are all heavily correlated so we only need to control for socioeconomic factor. We have yet to determine which socioeconomic factor we will use.

The current models that we have built have not looked extensively at these controls. Using WEKA, a free open-source data mining software, we have built a linear regression model that utilizes the binary values of the presence of a law (0 - law is not present, 1 - law is present) as inputs and gives the predicted age-adjusted rate of gun violence as an output. We arrived at a promising correlation coefficient of 0.87, although we are under the impression that this value does not necessarily imply that this model will be great at identifying key laws. The model follows the simple format a1x1 + a2x2 + … + anxn + b, where “a” represents the weight of a law and “x” represents the binary input of the law. If we focus on one law and keep all the other inputs constant, as we change that law between a 0 and a 1 the predicted rate of gun violence would only change by an amount equal to the weight of that law. Therefore, we figured that the laws with weights of a high absolute value would be significant to look at. The laws with the most predictive regression coefficients are attached.

Analyzing these laws revealed some nonsensical results. For example, the model predicts that applying state background checks on all firearms will significantly reduce gun violence, but applying state background checks on handguns only will significantly increase gun violence. We realized that our data was probably not ideal for regression, although we don’t want to scrap the idea of linear regression altogether. Instead, we started sorting through the codebook of laws and realized that the majority of laws in the codebook are linked. In other words, they are coded via conditional statements such as “if dealer is coded as a 1, then dealerh is automatically coded as a 1.” This means that some combinations of laws are not possible. Dealer cannot be coded as a 0 if dealerh is coded a 1. These conditional statements inspired us to explore decision trees, which we hypothesize can better map out the restricted combinations of laws.

Next, we created a random forest model in WEKA and on our own via python. The random forest that we created in python proved to be more useful. We used a method from scikit that gave us feature importance scores. The feature importance score of a variable is the percentage of variance in the outputs caused by that one variable. We obtained a list of the top ten laws with the highest feature importance scores, which is also attached. Interestingly enough, none of the ten laws with high feature importance scores also have high magnitude regression coefficients.

That is a summary of our work so far. We plan to continue to pursue random forests and linear regression. We want to map out trees for all of the intercorrelated laws. An example tree is attached. We need to gather data for controls. We also need to figure out a concrete way to incorporate controls into our models. Currently, we are playing around with using controls as additional predictors, but we know that there are different ways to control for these factors. Again, our ultimate goal with this project is to be able to use these machine learning models to make suggestions for policymakers and clearly state that certain laws are beneficial and certain laws are not.

We will gratefully accept any insight or advice that you may have for us! It would be an incredible opportunity to work with you and we hope that you will be able to assist us and help move our project forward. We would like to set up a meeting with you via skype. The week after next is ideal for us. February 13th, and 15th work at any time between 9:15 and 10:55 or in the afternoon. Let us know when you can skype.

Please contact us with questions, comments, or concerns at [chrisgoodhart817@gmail.com](mailto:chrisgoodhart817@gmail.com) and [aryan.deorah@gmail.com](mailto:aryan.deorah@gmail.com).

Sincerely,  
  
Chris Goodhart and Aryan Deorah