2020 Fall CS534 Final Project Proposal

Howard Huang, Sean O’Connell, and Sagar Soni

# Introduction – Sean

What school spending categories correspond to violent crime rate?

For the purposes for this study, we are focusing on Georgia counties.

# Data Exploration – Sean

# Past Work – Howard

# Evaluation and Propose Technique – Sagar

We plan to use the data from the U.S. Census survey data for educational spending by district by year. This will then be used to find patterns within a time-delayed crime rate data for districts in Georgia. The aim is to predict the level of crime in a specific year due to the categorical education spending by a specific district. Since we have more features than we do data points due to the various types and amounts of spend and transformation that are possible for the feature compared to crime data being reported from 19XX to 20XX. Thus, we will be using a Lasso based linear regression on the base feature sets to set a baseline for the most important spend categories that is able to predict the crime rate in a given district for a given time delay. Once the accuracy, defined by R square and MSE, is benchmarked for a specific feature set, we will be conducting dimensional reduction to see if there are spending in multiple dimensions that can correlate to the crime rate target variable more accurately. We believe that in general, while 1 specific type of educational spending will not lower crime rate. A mixture of spending across various categories that are linked (i.e. Teaching salary and yearly classroom supplies) and can show a correlation between increased spending and decreased crime rates.

To normalize the impact of population changes impacting the crime rate, we will be determining the crime rate per 1000 people. Additionally, the education spend will be normalized against the nominal GDP for Georgia which should eliminate some of the variability in the changing demographics of Georgians across the timeframe that we are observing.

While the changing demograpics is a concern for the time series data. The data related concern is that for the limited data, we could face an overfitting scenario with many of the techniques we will try due to the features >> data set.

A success for the model will be to explain in *plain English* where and how districts should spend their education budget on to decrease crime rates in a specified time delay.s

Linear Regression

Lasso for individual

Dimensionality reduction for combined features

MSE ->

# References