School Crime and Safety

Katherine Lee and Mark Rauschkolb

December 19, 2021

Contents

1 Executive Summary

Problem. Sandy Hook Elementary School and Santa Fe High School are synonymous with school shootings and violence. While national trends show that violence and victimization has declined from 1992 to 2017, there is a sentiment that the number of school shootings is increasing, mainly due to a steep increase in multiple-victim homicides (as opposed to single-victim homicides), revealing that the topic of school violence is still very relevant. On top of this, smaller instances of school crime are often forgotten even though they are still abundant and extremely detrimental to educational outcomes and children safety. For our final project, we decided to look into which variables were predictive of crimes and incidents of violence across US schools throughout 2015. While only analyzing 2015 school trends inherently limits the generalizability of our results, we believe the analysis draws high-level analyses about school crime trends.

Data. We draw from two different types of datasets: school-level data and county-level data. We used school-level data from the Urban Institute Education Portal to understand what school-specific features may be predictive of school crime. Using this portal, we were able to access data from: National Center for Education Statistics' Common Core of Data (CCD), the Civil Rights Data Collection (CRDC), the US Department of Education's EDFacts, and IPUMS' National Historical Geographic Information System (NHGIS). Specifically, we pulled: school-level directory / geographic data, school finance data, teacher and staff data, discipline instances, and crime incidences. We used county-level data to analyze school crime macro-trends. This county-level data came from US Economic Research Service, US Quarterly Census of Employment and wages, and Census Bureau's Demographic data. Specifically, we pulled: unemployment, education, poverty, wage, and demographic data. Our response variable was total crime per 1,000 students enrolled. We created this variable by adding a variety of crime incidents (rape, assault, robbery, etc.) and dividing it by the schools' respective student enrollment.

Analysis. Before running any analyses, we split our data into a training dataset and a testing dataset. Because of the heavily-skewed nature of our response variable, total crime incidents per 1,000 students, we square-root transformed the response, which resulted in more symmetric data. Then, on our train data, we explored our data to understand correlations between the features (through correlation plots) and between the features and the response. To understand the predictors of total crime and make a predictive model, we built five different cross-validated models, with the end goal of selecting the best model with the best performance. Specifically, we built a ridge regression, Lasso regression, pruned regression decision tree, random forest, and boosting model. Both penalized regression models gave similar RMSEs, with the Lasso method yielding slightly better predictions. Of the tree-based models, the random forest and boosting model greatly outperformed the regular regression decision tree. The Random Forest had the lowest test error of the tree-based models, and served as the best predictive model overall.

Conclusions. Throughout all models, we consistently saw similar variables surface as highly predictive of total crime. In general, we saw FINISH

2 Introduction

Background. School violence has always been a part of the United States' education system. From 2017-2018, 80% of public schools recorded one or more incidents of violence, translating to a crime rate of 29 incidents per 1,000 students enrolled.[1] Not all recorded incidents of violence and crimes were reported to the police, suggesting that the crime rate in America's schools is even higher. While the total victimization rate and the rates of specific crimes have declined since 1992, crime rates are still high, suggesting it is very important to utilize various data sources to understand the factors that contribute to school crime. Furthermore, a thorough analysis of school crime rates and predictive factors may help inform strategies to improve school safety precautions. Establishing reliable indicators of school crime and safety are important in ensuring the safety of schools across America.

Prior research and articles have shown that school crime are influenced by a variety of factors. For example, research has shown that higher use of police in schools is associated with more school crimes.[2] Moreover, trends show that students residing in rural areas had higher rates of total victimization than students residing in suburban areas, suggesting crime rate differences across school geographies.[3] Despite past research, there is much more to be learned and there is insufficient research regarding the relationship between school crime rates, school-specific features, and macroeconomic indicators.

Analysis goals. Given school crime can be attributed to a variety of factors, including both school-specific variables and demographic/county-level variables, we explored how total crime rates (total crime per 1000 enrolled students) are affected by various school and demographic features. Specifically, we were interested in what kind of factors, either school-specific or county-specific, and which specific variables are most predictive in school crime rates. Some examples of school-specific features that are included in our dataset are: school type, whether the school is a charter school, how many security guards are employed, and the average teacher salary. Some examples of county-level features that we included are: the percentage of the county population that attended college, the proportion of the population that are white, and labor force participation. We built a series of regression models and tree-based models to predict school crime rates, given school-specific features and county-level features, on our training dataset, and evaluated the success of each model by comparing the model predictions to actual observed rates in our testing dataset (using RMSE).

Significance. Our analysis will contribute to research regarding school crime risk factors and inform government officials or school administrators what factors are predictive of school crime. This will help support efforts to minimize school crime and improve school safety. For instance, we believe our findings can help schools allocate budgets more effectively, create systems to track possible increases in crime before they happen, or even give local goverments some direction with regard to which macroeconomic factors they should begin targeting to reduce school crime rates.

[1] School Crime: Fast Facts. (n.d.). https://nces.ed.gov/fastfacts/display.asp?id=49.

[2] Gottfredson, D., Na, C., (2011). Police Officers in Schools: Effects on School Crime and the Processing of Offending Behaviors

[3] National Center for Education Statistics. (n.d.). Indicators of School Crime and Safety: 2015. https://bjs.ojp.gov/content/pub/pdf/iscs15.pdf

3 Data

3.1 Data sources

Our dataset was merged from data from the following sources: 1) Urban Institute Education Portal, 2) US Economic Research Service, 3) US Quarterly Census of Employment and wages, and 4) Census Bureau's Demographic data. Each data source includes multiple years of data, and we chose to focus on 2015 data since there were the most number of available features for that year.

The data regarding school-specific features came from the Urban Institute Education Portal. The data from the portal was pulled from a variety of sources, including the National Center for Education Statistics' CCD, the CRDC, the US Department of Education's EDFacts, and IPUMS' NHGIS. In order to ensure a wide range of explanatory variables in our dataset, we pulled 2015 variables across the following categories: school directory, demographic, finance, teacher and staff, and discipline data, in addition to our response variable, which came from the criminal data. The data was specially pulled using the "educationdata" library provided. After pulling these datasets, we merged the variables into a final dataset by each school's National Center for Education Statistics (NCES) identification number (a unique school identifier).

Because we were also interested in broader demographic data, we drew from various government and census datasets, including: US Economic Research Service, US Quarterly Census of Employment and Wages, and Census Bureau's Demographic data. The US Economic Research Service provided county-level unemployment, educational attainment, and poverty rates The unemployment and educational attainment rates were recorded from 2000 to 2020, and the only poverty estimates available were from 2019. Because the rest of our data is of 2015, we assume that poverty estimates from 2015 are reasonably similar to those from 2019. We extracted data from 2015 in the unemployment and educational attainment data and the 2019 poverty estimates for our data analysis. The US Quarterly Census of Employment and Wages had average wages for each county since 1990 - we selected 2015 directly from the dashboard's UI. Lastly, the Census Bureau's demographic data had total populations, as well as different race populations for each county from 2010 to 2019. Again, we filtered for only 2015. We merged the variables into a final dataset by each county's unique FIPS code.

3.2 Data cleaning

For data cleaning, we started by examining each dataset in depth to determine which features and observations to keep.

In the school-specific datasets, we first transformed all negative values to NA values, since negative values imply that the data was missing/not reported, not applicable or not interpretable, per the Education Data Portal. Then we dropped all features that had more than 85% `NA` values, and dropped repeat observations and observations with any `NA` feature values. We dropped `NA` values for consistency purposes, as many of the data mining methods we employed require that all variables be populated with non-NA fields. In the directory dataset, one important feature is the degree of urbanization (urban-centric locale) of the school, which was coded with 12 levels, so we collapsed the categorical variable into 4 broader levels, including city, suburb, town, and rural. After cleaning each individual dataset, we merged the school-specific sets by the schools' unique NCES identification number.

For the macro-level county datasets, we first filtered the data to include only 2015 data (except poverty estimates because 2015 data was not available). We renamed variables for clarity and computed various race metrics as a percentage of population for comparability purposes. Each school is associated with a county-specific FIPS code and the macro-level county data includes observations for each county, which allowed us to merge in the macro-level county data in with the school-specific data, resulting in our final dataset that we used for analysis.

3.3 Data description

3.3.1 Observations

Our cleaned and merged final dataset has a total of 66,694 observations, corresponding to each of the schools included in our analysis.

3.3.2 Response Variable

Our response variable is the number of crimes per 1000 students enrolled (square root transformed). We created our response variable by summing rape incidents, sexual battery incidents, robberies and attacks to get a high level variable that captured crime rates. We also transformed the total crime figure to be per 1000 students enrolled (by dividing by the school's enrollment and then multiplying by 1000), to improve comparability of schools on a 'per student enrolled' basis as it inherently controls for school size variability. Then we square root transformed the response to obtain a more "symmetric" variable.

3.3.3 Features

We included 45 explanatory variables in our analysis, which fall into two broad categories: school-specific factors and macro/county-level factors. For a detailed specification of these variables, refer to Appendix A.

3.4 Data allocation

After data cleaning, we split our dataset into a training dataset and a testing dataset. We used an 80-20 split, such that the training set consists of 80% of the observations and the testing set consists of the other 20% of the observations. The training dataset was used for building our predictive models and the testing dataset was used for model evaluation.

3.5 Data exploration

3.5.1 Response

We first explored the response variable's distribution. As seen in the histogram of total crimes per 1000 students variable (left plot of Figure 1). The histogram is very right-skewed, indicating that the dataset contains a number of outliers with extremely high crime rates. The mean is 19.3 crimes per 1000 enrolled students. To account for this, we square root transformed the response variable, which resulted in a distribution shown in the right plot of Figure 1. For our analyses and model building, we used the transformed response.

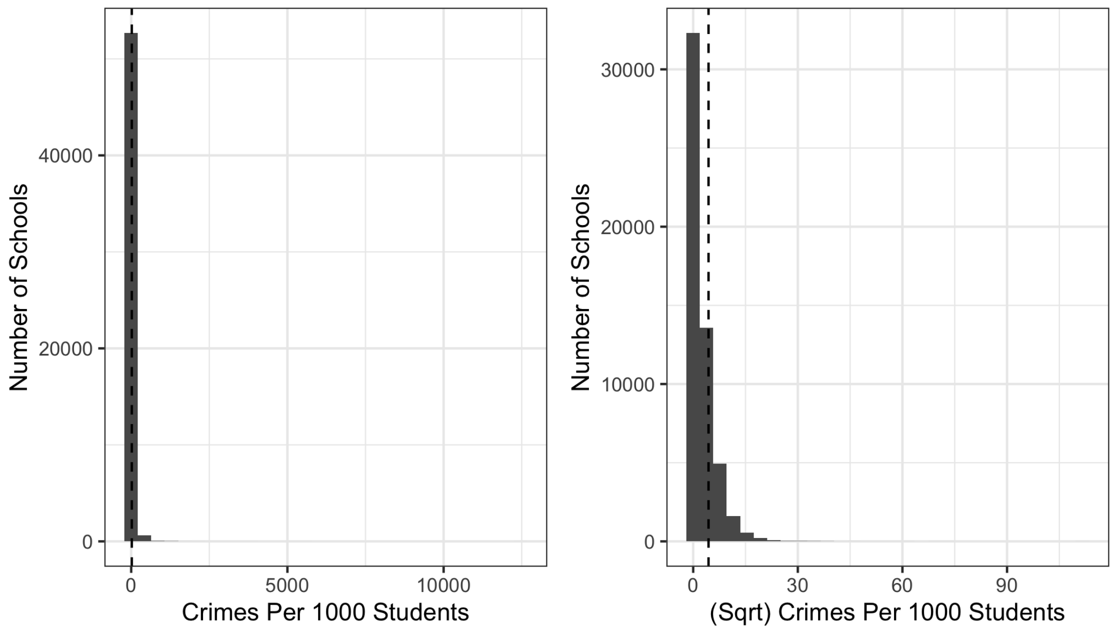

Figure 1: Distribution of Crime Rate (vertical dashed line indicates the mean)

We proceeded to determine which schools had extreme response rates by looking at the sorted data. The sorted data (Table 1) shows that Maple Lane School had the highest crime rate by far, more than double the second highest ranked school. Further research shows that Maple Lane School was a juvenile corrections facility that closed in 2010, indicating that the observation should not be in the dataset to begin with.[4] To account for this, we dropped Maple Lane School from the training set.

Table 1: Top Ten Schools By Crime Rate

School	Crimes per 1000 Students
Maple Lane School	12458
Lighthouse Academy - St. Johns	5407
LITTLE SCHOOL	4967
Allendale School	3956
Westbank Community School	3414
Carver Middle School	2990
Tangipahoa Alternative Solutions Program	2400
Monroe Area High School	2103
EASTERN WRIGHT PROGRAM	2000
Murrell School	1709

3.5.2 Features

Next, we explored high-level relationships and correlations of the predictor variables with the response variable. We first looked at correlations between a few school-specific features, as shown in Figure 2. We observe a positive correlation between the number of full-time equivalent teachers and the number of full-time equivalent security guards as well as teacher salaries. We also a negative correlation between the number of

[4] Turn Maple Lane into prison center? Allen, Marqise. https://www.theolympian.com/news/local/article25274248.html

full-time equivalent teachers and the number of students enrolled in free or reduced lunch. Moreover, there is a negative correlation between teacher salaries and the number of students enrolled in free or reduced lunch.

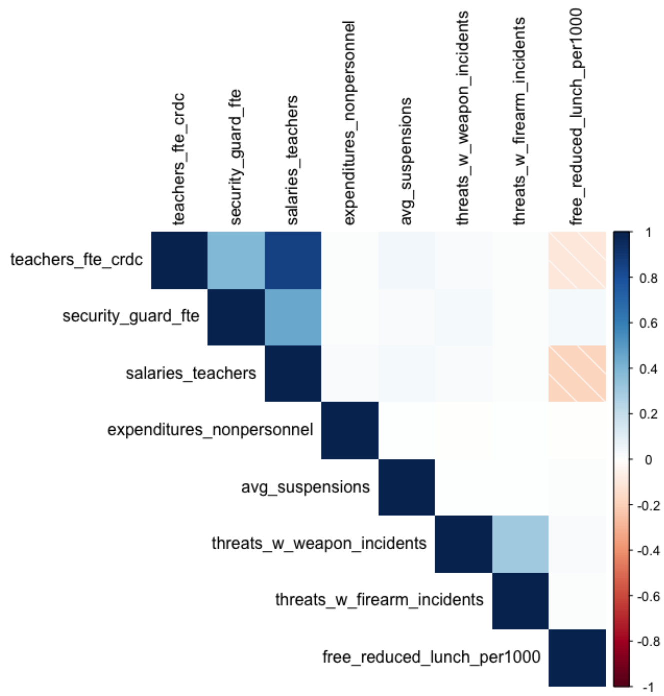

Figure 2: School-Specific Features Correlation Plot

We also looked at correlations between some county-level features, as shown in Figure 3. In general, county-level data exhibited much more multicollinearity. Many of these variables are directly related (the rates of college completion is negatively correlated with the rates of individuals who only completed high school). Other variables had obvious relationships, such as the rate of individuals who did not complete high school correlating negatively with household income and positively with unemployment; or rates of college completion correlating with household income. Another notable relationship was the negative correlation between white and black populations, which demonstrates that, to this day, American counties are still extremely segregated.

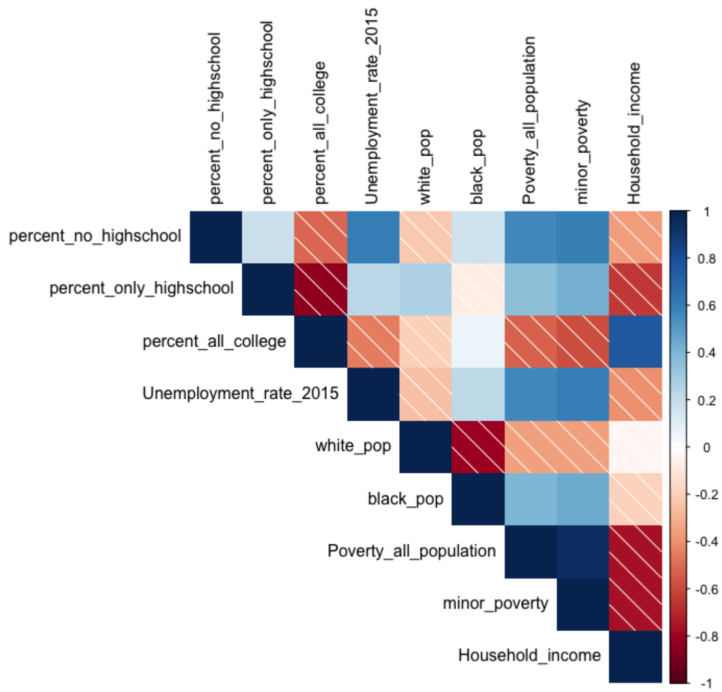

Figure 3: County-Level Features Correlation Plot

We also examined correlations across the different categories of variables (both school-specific and county-level) as shown in Figure 4. Once again, we observe a number of expected correlations (i.e. high rates of minors living in poverty is heavily correlated the rate of students enrolled in free/reduced lunch programs). Moreover, some relationships between a county's economic data and school resources began to emerge (i.e. higher poverty rates are correlated with lower teacher salaries). It was interesting to see that suspensions and number of threats did not correlate with any of the county-level socio-demographic factors. This analysis shows that the county-level macro data does have a direct impact on some of the resources of a school (i.e. teacher salary) but not necessarily on certain crime or disciplinary factors.

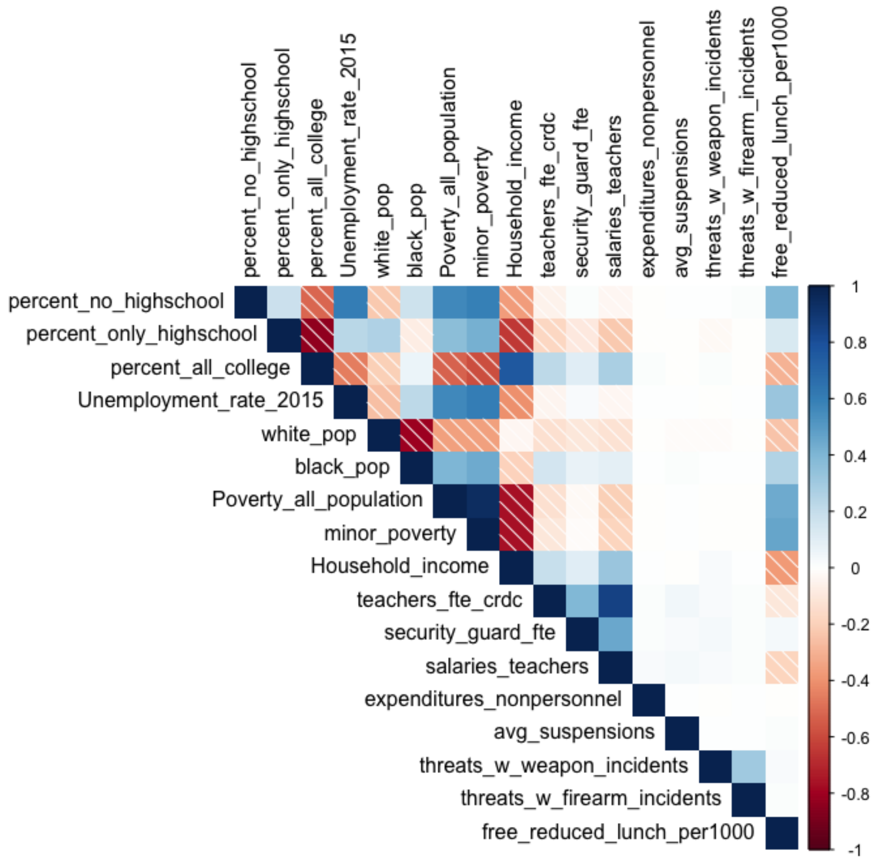

Figure 4: School-Specific and County-Level Features Correlation Plot

For further analysis, we explored how three variables (total poverty, unemployment, and threats) related to our response variable (total school crime) as an initial indicator to possible models. We did this by plotting scatterplots with our train data, along with a linear regression line to further illustrate the relationship, as seen in Figure 5. By inspection, threats with firearms or explosives, poverty rates, and unemployment rates were all positively correlated with crime rates. However, the relationship between threats and crimes is much stronger than the relationship between poverty / unemployment and crimes.

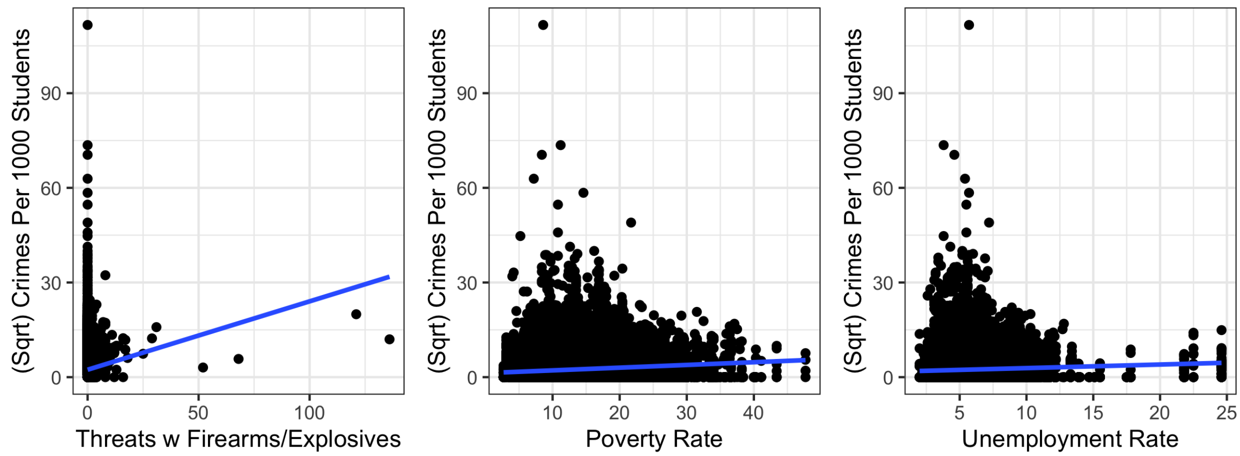

Figure 5: Crimes and Poverty Scatter Plot

We also explored how school type was related to school crimes. By inspection of Figure 6, there are significant outliers for each of the three types of schools, but the maximum for special education schools is the highest. The maximum crime rate out of all vocational schools was the lowest.

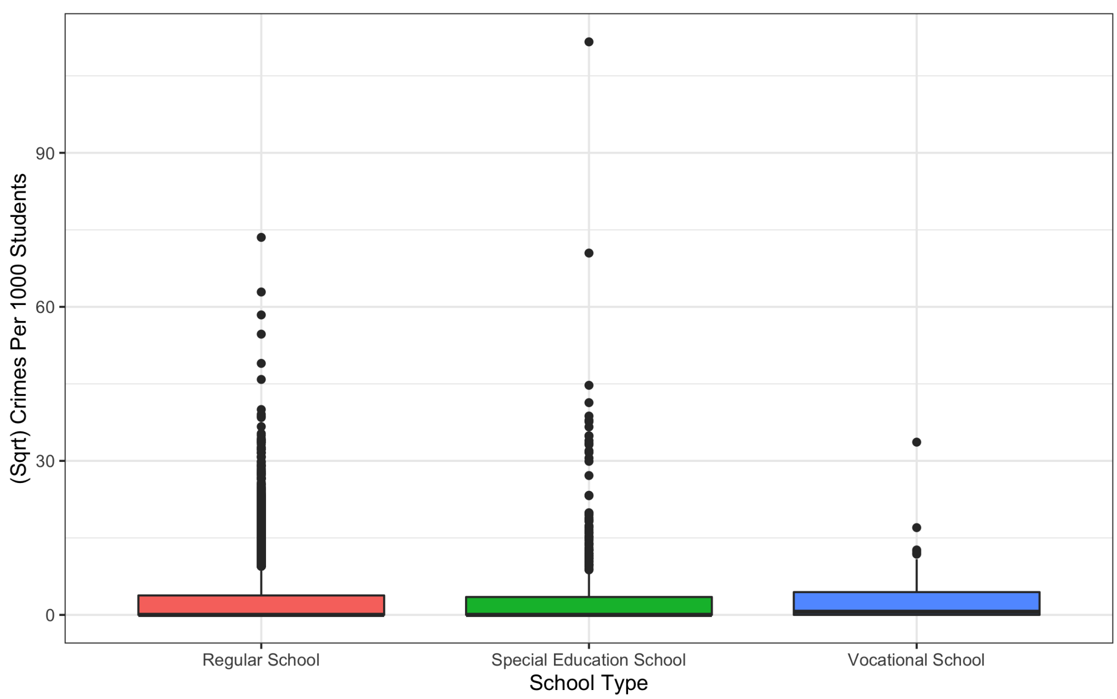

Figure 6: Crimes and School Type Box Plot

4 Modeling

4.1 Regression-based methods

4.1.1 Ordinary least squares

As a starting point, we built an ordinary least squares regression with all 45 explanatory variables. However, when considering the normality assumptions, we note that our response distribution, even after applying a square root transformation, is not normally distributed. Moreover, as demonstrated in our correlation plots, many of the features are correlated. Extended results are reported in the Appendix B (Figure 15). The R-squared of the regression was 0.144, signifying a rather weak relationship between the model and the response. We believe that severe skew in our data and multicollinearity between our predictor variables decreased the accuracy of our model.

However, there were a number of significant variables, including: the urbanization of the school location, a variety of quantity teacher metrics, the amounts of threats each school was reported with, race demographic data, and almost all economic data (unemployment, poverty, weekly wage, etc.).

4.1.2 Penalized regressions

Despite identifying significant variables from the ordinary least squares regression, the method utilized all of 45 explanatory variables (with high intercorrelations), which could lead to a cost in variance, and thus inaccurate predictions. To combat these issues, we built and evaluated shrinkage models, namely ridge regression and lasso regression, with the goal of fitting a more parsimonious and interpretable model.[5] Specifically, ridge regression is more stable when handling correlated features, as it "splits the credit" among correlated features, and lasso regression penalizes many features to 0, contributing to increased interpretability. For both penalized regression methods, we ran a 10-fold cross validation to optimize the choice of regularization parameters (λ).

The lasso regression trace plot is shown in Figure 7 and the selected features and respective coefficients are displayed in Table 2. We applied the one standard error rule to select the optimal λ value, and we notice that the lasso regression selects three variables: the number of students receiving free or reduced lunches per 1000 enrolled students, the proportion of the county population that identifies as black, and the amount of non-weapon threats a school receives.

Table 2: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
free_reduced_lunch_per1000	0.35
black_pop	0.07
threats_no_weapon_incidents	0.05

[5]Note: we did not display elastic net results because the elastic net selected an α value of 1, which reduces to the lasso regression

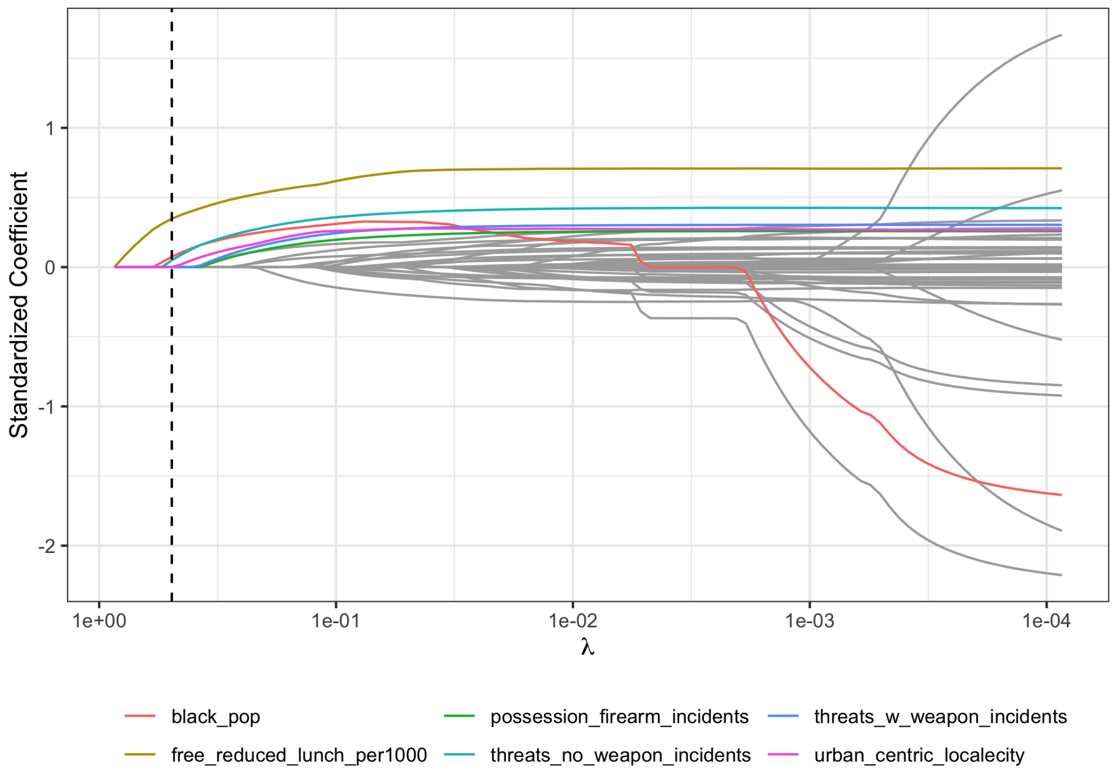

Figure 7: Lasso Regression Trace Plot

[why is this here] Following, in 3 are the results for tree based methods:

Table 3: RMSE summary for regression methods

Method	Test RMSE (transformed)	Test RMSE
OLS	3.39	62.0
Intercept-Only	3.65	60.4
Ridge	3.65	60.4
Lasso	3.60	60.1

4.2 Tree-Based Methods

4.2.1 Ordinary pruned tree

In addition to regression methods, we also implemented tree-based methods to capture possible non-linear relationships between the features and the response. We began our tree-based modeling by prediction via a traditional decision tree, which offers the benefit of high interpretability. We selected the complexity of the decision tree based on cost complexity pruning and cross-validation. The cross-validation plot compared to the number of terminal nodes is displayed in Figure 8. Based on the one-standard-error rule, we chose 6 terminal nodes.

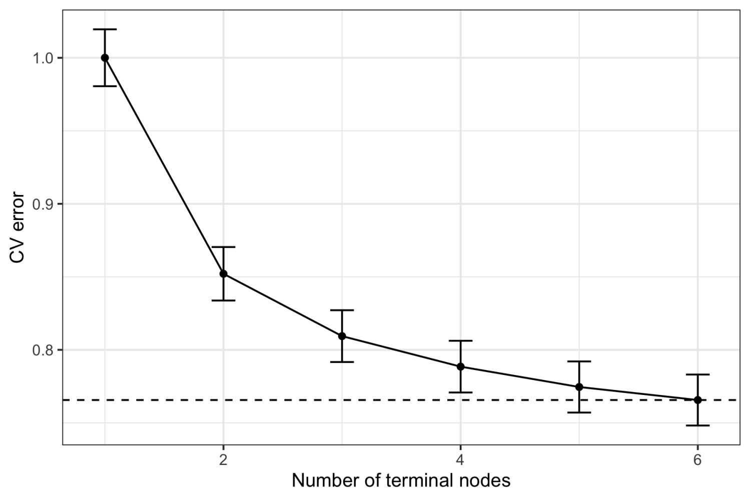

Figure 8: Decision Tree: Cross Validation for Terminal Nodes

To obtain the optimal pruned tree, we specified the chosen complexity parameter and applied cost complexity pruning. This results in the decision tree displayed in Figure 9. The path that leads to the greatest prediction of the square root transformed total crimes per 1000 enrolled students response is: the school experienced 1 or more no-weapon threats, more than 24 average student suspensions, and more than 638 students out of 1000 enrolled students receiving free or reduced price lunch. Logically, this path makes sense as more threats and more student suspensions are typically associated with unsafe school conditions, and more students enrolled in free or reduced priced lunches indicates a poorer county, which can be subject to higher crime rates.

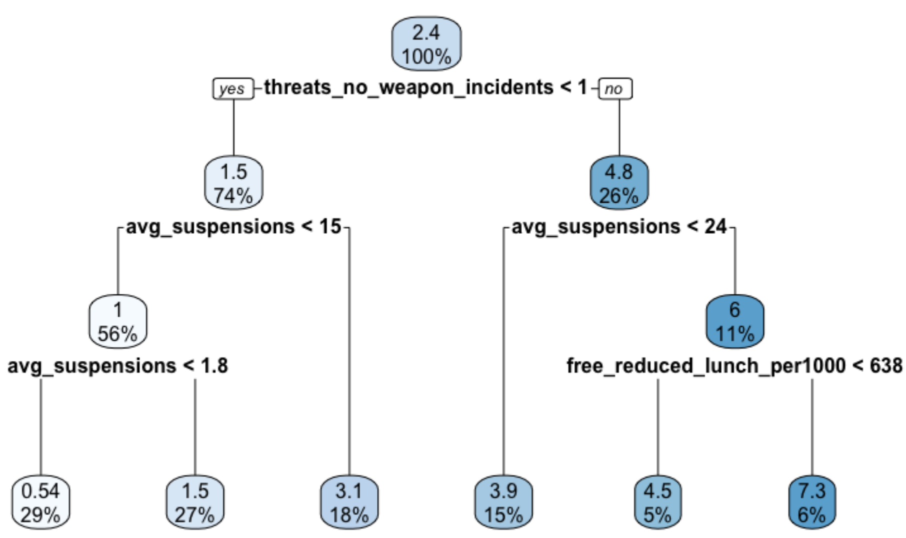

Figure 9: Pruned Decision Tree

4.2.2 Random forest

In order to improve the predictions from the individual tree, we proceeded to fit a random forest, which involves growing deep decision trees in parallel. Bagging is achieved when considering all 45 features at each split point of the tree, which leads to high variance and inaccurate prediction performance. To account for this, when tuning the random forest, we optimized both the number of bootstrap samples, corresponding to the number of trees, (n), and the number of variables to sample at each split point (m_{try}). Because of the computation costs of random forests and computer issues, we trained it on 5,000 observations of our train data.

First we trained a random forest with default settings, and visualized the fit displayed in Figure 10, which plots the OOB error as a function of the number of trees. By inspection, the error flattens out as soon as the number of trees is large enough (in this case, the error stabilizes around 200 trees).

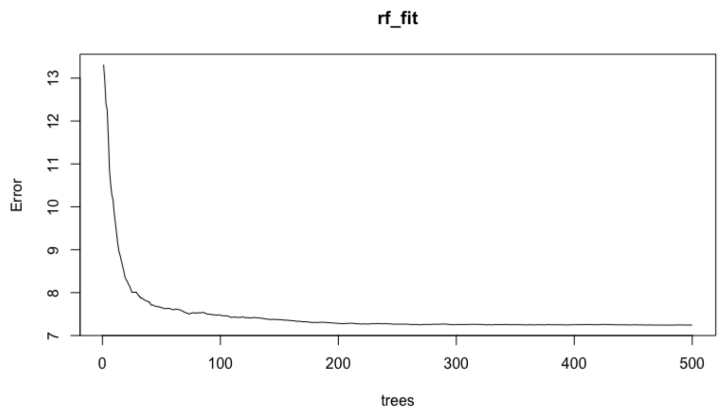

Figure 10: Default Random Forest OOB Error vs. Number of Trees

For the purpose of tuning m_{try}, we select 500 trees. We tune m_{try} using a systematic approach by choosing a grid of values of m_{try} and plotting the OOB error for 500 trees. Specifically, we trained the model using m_{try} values ranging from 1 to 30. The OOB error for each value of m_{try} is visualized in Figure 11. By observation, the OOB error is minimized at a value of $m_{try} = 11$. After reaching a minimum at this value, the OOB error begins to exhibit overfitting as it begins to trend upward.

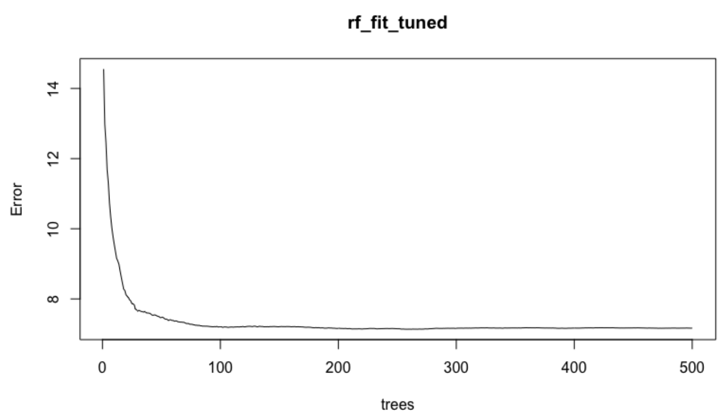

Figure 11: Tuned Random Forest OOB (m = 11)

Because we do not explicitly tune n, or the number of trees, we simply select the largest value, which is $n = 500$. Then, we fit the random forest using $n = 500$ and $m_{try} = 11$. We then plotted the variable importance using the two types of variable importance, purity-based and OOB-based variable importance. The results from both of these variable importance measures is given in Figure 12. By inspection, we observe that the number of no-weapon threats and the number of average suspensions of a school have the highest importance as measured by both the purity-based and OOB-based metrics. The proportion of the population that identifies as black, the proportion of the population that identifies as white, and the number of teachers in the previous school year are also variables that are important. This suggests that these variables are the most important in predicting total crimes across different schools in America.

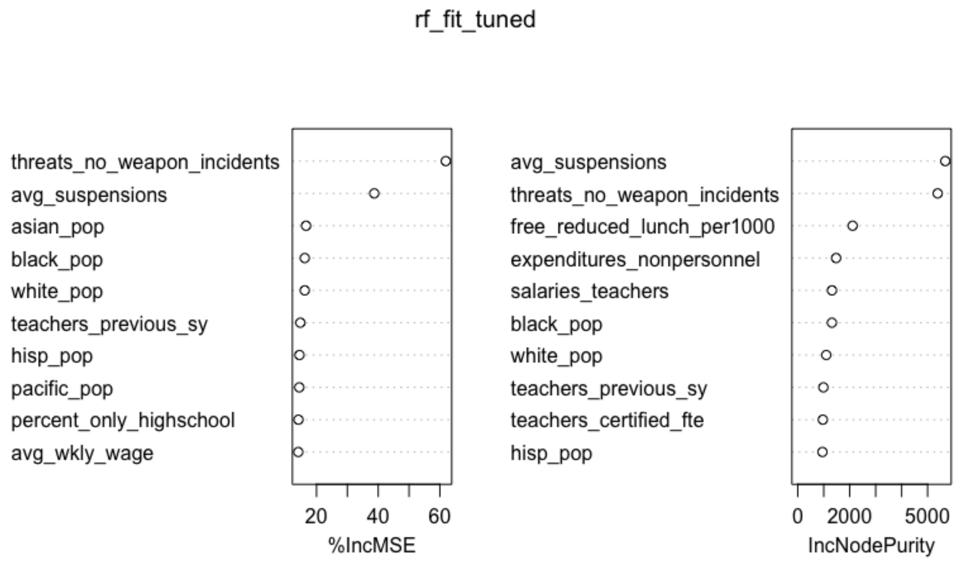

Figure 12: Tuned Random Forest Variable Importance

4.2.3 Boosting

In addition to the random forest model, we also implemented a gradient boosting model, which is another method of aggregating multiple decision trees to improve prediction performance over a traditional decision tree. Boosting grows shallow decision trees sequentially, by considering a low-complexity weak learner (a shallow decision tree) and boosting the performance of the weak learning by applying an iterative method. For our first boosting model, we used default parameters (500 trees, an interaction depth of 1, and a shrinkage of 0.1). To tune this model, we adapted both the depth of each individual ‘weak-learner’ (interaction depth) and the number of ‘weak-learners’ (number of trees) used per model. After noticing that using a number of trees between 500 and 1000 was optimal, we tuned the interaction depth by testing interaction depth levels of 1, 2, and 3. By inspection of the cross-validation plot shown in Figure 13, we see that an interaction depth of 3 offers the lowest cross-validation error. Specifically, the cross-validation error reaches its minimum with 518 trees.

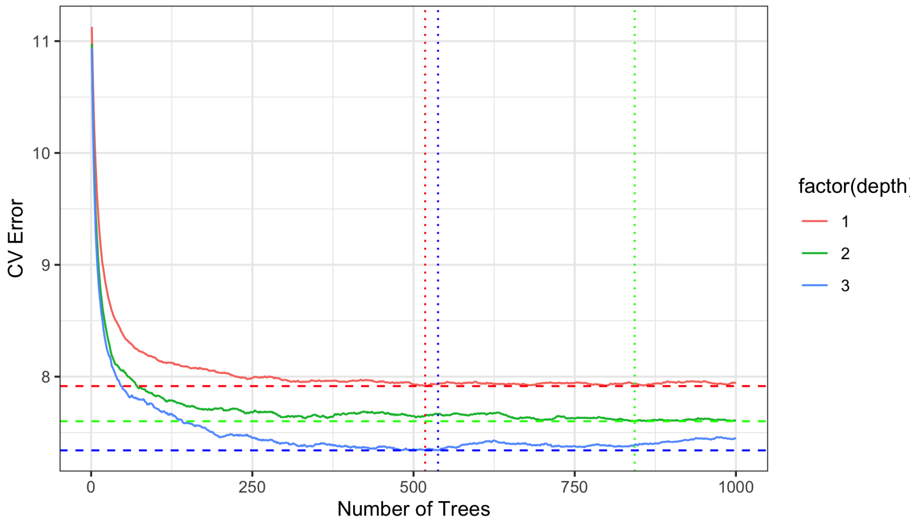

Figure 13: Boosting Tuning Interaction Depth and Number of Trees

With the tuned boosting model using 518 trees and an interaction depth of 3, we judged variable importance by comparing purity-based importance. The ranking of variables based on purity-based importance is given in Table 4. Similar to the random forest method, the number of suspensions and the number of no-weapon threats rank high in terms of variable importance.

Table 4: Boosting top 10 most important features based on purity score

Feature	Coefficient
avg_suspensions	18.43
threats_no_weapon_incidents	15.53
free_reduced_lunch_per1000	6.20
expenditures_nonpersonnel	3.81
salaries_teachers	3.55
black_pop	3.34
white_pop	2.90
labor_force_participation	2.74
teachers_previous_sy	2.58
pacific_pop	2.40

In addition, we also examined partial dependence plots of average suspensions and the number of no-weapon threats, which are the two most important features, as shown in Figure 14. It is important to remember that because our tuned boosting model uses an interaction depth of 3, they are simply an approximation for illustrative purposes, but they nonetheless give a good representation of how each variable affects total crime. Taken together, Figure 14 suggests that schools that suspend more students and face more no-weapon threats face disproportionately higher crime rates.s

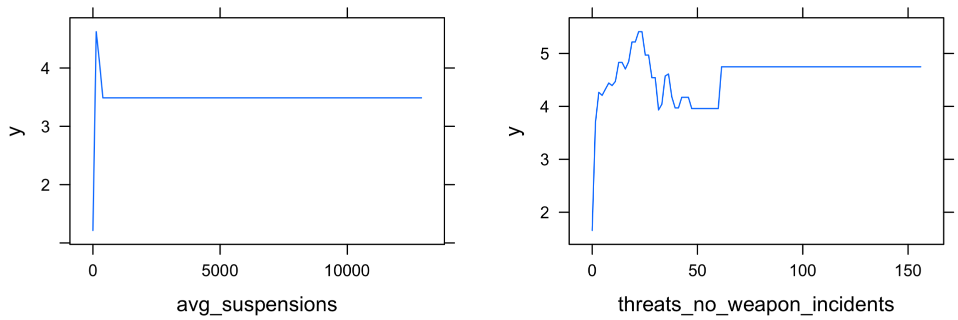

Figure 14: Average Suspensions and No-Weapon Threats Partial Dependence Plots

5 Conclusions

5.1 Method comparison

Table 5: RMSE Summary for All Methods

Method	RMSE
OLS	62.0
Intercept-Only	60.4
Ridge	60.4
Lasso	60.1
Tree	57.0
Random Forest	53.8
Boosting	52.3

Table 5 displays the test RMSE (untransformed) for all the methods considered. Based on RMSE, the random forest and boosting methods have the lowest test errors. This is reasonable given these models' tendencies to have high predictive accuracy. Between the two, the boosting model has the lowest test error of 52.3 crimes per 1000 students, but its accuracy is comparable to the random forest method, which has a test error of 53.8 crimes per 1000 students. Notably, the tree-based methods perform better than the regression-based methods, namely OLS, a predictor that simply predicts the mean crime rate of the training set, ridge regression, and lasso regression. This ultimately points to the idea that the relationship between the response and the explanatory variables is not necessarily linear, explaining why the tree-based methods outperformed the regression-based methods.

Regardless of the test errors, all of the models overlap when considering important features. For example,

Regardless of the magnitudes of the test RMSEs, of these differences in test MSE, the methods overlap significantly in their identification of important variables from the larger set. For instance, the elastic net regression selects the following variables, which are also selected by LASSO and deemed significant in the OLS model: other providers ratio, unemployment, income inequality, housing overcrowding, residential segregation—non-White/White, homeownership, and physical inactivity. The random forest and boosting models both include low birthweight percentage, median income, and unemployment percentage in the top 10 most important variables, as measured by their contributions to node purity.

5.2 Takeaways

1. As seen by our regression tree, random forest, and boosting model, school-specific features were more important predictors than county-level features. Although many macro variables (unemployment, poverty, low adult education) did correlate with school crime, these factors became less relevant when accounting for school-specific variables. I believe this is a sign that although schools may be placed in a more impoverished county, good funding, staff etc. in a school can help the school beat its precarious circumstances. Counselors and teachers can help create a safe environment for students irrelevant of the environment the school is placed in.

2. Lasso penalized regression revealed two county-level variables were % of a county that identifies as being black, and the % of 5-17 year old children in poverty. I think this signals two key aspects to local government: Firstly, all kinds data repeatedly shows that marginalized minorities are afflicted by a variety of issues the most; although it is already known, we believe local government needs to remember to recognize inequalities in race when allocating funds. Next, we believe it is important to track teenage poverty specifically - helping families that are sustaining younger children and adolescents could reduce crime and improve school systems.

3. How much a school is suspending its students is the most important variable in our regression tree, random forest, and boosting models. We believe the department of education should investigate this phenomenon more. There may be reverse causation (a lot of crime results in higher suspension rates) however, after some research we found a many analyses that corroborate that suspensions may actually be causing worse behavior such as this National Association of School Psychologists article, or this National Education Association article, both of which explains how suspension creates a vicious cycle: antagonizing students, sending them away school to school which often lead to long-periods of unsupervised time, more crime upon returning to school, and a subsequent suspension. Reform on this traditional reprimand method could be a cost-free method of greatly reducing crime rates.

4. Investments in teachers and counselors seem to be the most important. Variables such as security guards, law enforcement, administration, and nonpersonnel expenditures did not prove to be important in any of our models. On the other hand, larger investment in teachers and counselors (those who directly interact with students) did show importance. I believe this is extremely important for school boards to consider when allocating budgets.

5.3 Limitations

5.3.1 Dataset limitations

As it is detailed in the Frequently Asked Questions page of County Health Rankings & Roadmaps program website,[6] all of the variables are from 2019 or earlier. Thus, it is possible that the values of the variables assessed were different in 2020, meaning that the interpretation of our analysis may need to be taken with a grain of salt. Regardless, given that we analyzed data on the county level, it is unlikely that any county experienced enough drastic change over the course of the year to significantly affect our analysis. Furthermore, our dataset has a large number of observations to account for potential variability, although notably, many observations had to be removed as some of the R packages used to build some of our models required that no NA values were present in the data. In other words, since each observation represents a different US county, many counties were left out of our analysis. Another limitation is that, as described in the Exploratory Data Analysis section, there is evidence of correlation amongst some of our explanatory variables. This means that some variables can be confounding variables, which mask or distort the relationship between measured variables. Also, variables selected in the LASSO regression and elastic net regression as well as variables marked important in the tree methods might be misleading in that, given how variable selection works, it is possible that some selected variables are simply representative of a larger group of correlated variables.

[6] County Health Rankings & Roadmaps. (n.d.) Frequently Asked Questions. https://www.countyhealthrankings.org/explore-health-rankings/faq-page

5.3.2 Analysis limitations

While splitting the data into training and testing datasets allows for a more unbiased test of the models, we recognize that our conclusions inherently contain some randomness due to the random split of the data. In other words, splitting the data again using a different random seed may have yielded different p-values in the OLS regression, different selected variables in the shrinkage methods, and different variables selected as important in the tree-based methods. Next, although we provide different methods for robust interpretation of the variables, our analysis incorporates only a specific subset of health related variables. The results of the analysis might change dramatically if we were to incorporate other variables. For example, as mentioned in the Exploratory Data Analysis, states in the northeastern part of the United States suffered from a high rate of COVID-19 cases and deaths in 2020. This may not have been because these states performed poorly in the health variables mentioned above. Rather, it may be due to the fact that these states are densely populated and hence were more susceptible to disease spread at the outset of the COVID-19 pandemic. In other words, other factors like geographic or demographic variables can hugely impact case fatality rate.

5.4 Follow-ups

To compensate for the limitations mentioned above, more extensive analysis can be done as we acquire more data from 2020 and 2021. Not only can we extend our analysis by utilizing the most up-to-date datasets, but we can also examine how COVID-19 cases and deaths have affected various health factors of each county. In other words, the explanatory and response variables can be reversed to conduct more dynamic data analyses. Next, given that many observations needed to be omitted in our dataset as they contained NA fields, we recommend that our analyses be reconducted once the missing data is collected. Finally, future work on the social determinants of health in the context of COVID-19 might also look at different population levels such as states, bigger geographical regions in America, or even different countries.

A Appendix: Descriptions of features

Below are the 45 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.

School-Specific Variables:

- *Geography*
  - Urbanization (`urban_centric_locale`): Factor variable representing the degree of urbanization. Factors include schools located in: cities, rural areas, suburbs, and towns.
- *School Attributes*
  - School type (`school_type`): Factor variable representing the school type. Factors include: regular schools, special education schools, and vocational schools.
  - Charter school (`charter`): Binary variable representing whether the school is a charter school (1) or not a charter school (0).
  - Students eligible for free or reduced-price lunch (`free_reduced_lunch_per1000`): Number of students eligible for free or reduced-price lunch per 1000 enrolled students.
- *Teacher and Staff*
  - Full-time equivalent teachers (Civil Rights Data Collection) (`teachers_fte_crdc`): Number of full-time equivalent teachers.
  - Full-time equivalent certified teachers (`teachers_certified_fte`): Number of full-time equivalent certified teachers.
  - Full-time equivalent uncertified teachers (`teachers_uncertified_fte`): Number of full-time equivalent uncertified teachers.
  - Full-time equivalent first-year teachers (`teachers_first_year_fte`): Number of full-time equivalent first-year teachers.
  - Full-time equivalent second-year teachers (`teachers_second_year_fte`): Number of full-time equivalent second-year teachers.
  - Current school year teachers (`teachers_current_sy`): Number of current school year teachers.

  - Previous school year teachers (`teachers_previous_sy`): Number of previous school year teachers.
  - Full-time equivalent teachers absent more than 10 school days (`teachers_absent_fte`): Number of full-time equivalent teachers absent more than 10 school days
  - Full-time equivalent school counselors (`counselors_fte`): Number of full-time equivalent school counselors.
  - Full-time equivalent psychologists (`psychologists_fte`): Number of full-time equivalent psychologists.
  - Full-time equivalent social workers (`social_workers_fte`): Number of full-time equivalent social workers.
  - Full-time equivalent nurses (`nurses_fte`): Number of full-time equivalent nurses.
  - Full-time equivalent security guards (`security_guard_fte`): Number of full-time equivalent security guards.
  - Sworn law enforcement officers indicator (`law_enforcement_ind`): Indicator of whether a sworn law enforcement officer has been assigned to the school.
  - Teacher salaries (`salaries_teachers`): Personnel salaries at school level (teachers only) amount.
- *School Finance*
  - Non-personnel expenditures (`expenditures_nonpersonnel`): Amount of non-personnel expenditures. May include: professional development for teachers, computers, library books, and other learning materials.

- *Criminal Activity*
  - Threats of physical attack with a weapon (`threats_w_weapon_incidents`): Number of incidents of threats of physical attack with a weapon.
  - Threats of physical attack with a firearm or explosive device (`threats_w_firearm_incidents`): Number of incidents of threats of physical attack with a firearm or explosive device.
  - Threats of physical attack without a weapon (`threats_no_weapon_incidents`): Number of incidents of threats of physical attack without a weapon.
  - Possession of a firearm or explosive device (`possession_firearm_incidents`): Number of incidents of possession of a firearm or explosive device.
- *School Discipline*
  - Suspensions (`avg_suspensions`): Number of student suspensions.

County-Level Variables: - *Education Variables* - Percent of adults without higschool (`perecent_no_highschool`): Percentage of adults that do not have a higschool diploma for a specific county - Percent of adults with highschool diploma but did not begin college (`percent_only_highschool`) for a specific county - Percent of adults with some college (`perecent_some_college`): Percentage of adults that have a highschool diploma, and started (but did not complete) college education for a specific county - Percent of adults with college (`perecent_all_college`): Percentage of adults that have a college degree for a specific county - *Employment Statistics* - Civilian labor force (`Civilian_labor_force_2015`): How many civilians are working in the labor force (both employed and unemployed). - Unemployment rate (`Unemployment_rate_2015`): the unemployment rate for a county - Labor force participation (`labor_force_participation`): (employed + unemployed) divided by population - *Demographics* - Population (`pop`): total county population - Percent white (`white_pop`): percentage of a county that identifies as white - Percent black (`black_pop`): percentage of a county that identifies as black - Percent asian (`asian_pop`): percentage of a county that identifies as asian - Percent indian (`indian_pop`): percentage of a county that identifies as indian - Percent hispanic (`hisp_pop`): percentage of a county that identifies as hispanic - *Poverty* - General poverty rate (`Poverty_all_population`):percentage of county living in poverty
- Minor poverty rate (`minor_poverty`): percentage of minors (0-18 years) in county living in poverty - School-age poverty rate (`percent_school_age_children_poverty`): percentage of school age children (5-18 years) in county living in poverty - *Income* - Household Income (`Household_income`): Average yearly annual income earned by a household - Weekly Wage (`avg_wkly_wage`): The average weekly wage paid by employers in a county

B Appendix: Other

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      -2.07e+01   2.65e+01   -0.78  0.43412    
urban_centric_localerural        -8.56e-01   5.17e-02  -16.56  < 2e-16 ***
urban_centric_localesuburb       -6.14e-01   4.31e-02  -14.26  < 2e-16 ***
urban_centric_localetown         -7.03e-01   5.92e-02  -11.86  < 2e-16 ***
school_type2                      9.07e-01   1.45e-01    6.25  4.1e-10 ***
school_type3                      7.90e-01   3.39e-01    2.33  0.01971 *  
charter1                         -4.45e-01   6.74e-02   -6.60  4.2e-11 ***
teachers_fte_crdc                -2.59e-01   4.10e-02   -6.33  2.5e-10 ***
teachers_certified_fte            2.61e-01   4.11e-02    6.37  1.9e-10 ***
teachers_uncertified_fte          2.71e-01   4.13e-02    6.57  4.9e-11 ***
teachers_first_year_fte           4.49e-02   5.43e-03    8.25  < 2e-16 ***
teachers_second_year_fte          5.37e-03   5.31e-03    1.01  0.31169    
teachers_current_sy              -5.74e-03   1.98e-03   -2.90  0.00371 ** 
teachers_previous_sy             -2.27e-03   9.34e-04   -2.43  0.01530 *  
teachers_absent_fte               2.42e-02   1.88e-03   12.87  < 2e-16 ***
counselors_fte                    4.27e-02   7.66e-03    5.57  2.5e-08 ***
psychologists_fte                -9.19e-03   1.02e-02   -0.90  0.36887    
social_workers_fte                1.84e-01   2.25e-02    8.18  3.0e-16 ***
nurses_fte                       -8.45e-04   3.61e-03   -0.23  0.81509    
security_guard_fte                1.60e-02   1.23e-02    1.30  0.19331    
law_enforcement_ind               4.82e-01   3.65e-02   13.21  < 2e-16 ***
salaries_teachers                -4.97e-08   1.98e-08   -2.51  0.01209 *  
expenditures_nonpersonnel        -1.24e-09   1.07e-09   -1.16  0.24561    
threats_w_weapon_incidents        2.19e-01   1.12e-02   19.51  < 2e-16 ***
threats_w_firearm_incidents       6.08e-02   1.59e-02    3.82  0.00013 ***
threats_no_weapon_incidents       2.54e-02   8.99e-04   28.27  < 2e-16 ***
possession_firearm_incidents      4.67e-01   2.73e-02   17.07  < 2e-16 ***
avg_suspensions                   9.34e-06   1.36e-05    0.69  0.49200    
free_reduced_lunch_per1000        2.60e-03   6.59e-05   39.37  < 2e-16 ***
percent_no_highschool             3.32e-01   2.65e-01    1.25  0.21038    
percent_only_highschool           2.89e-01   2.65e-01    1.09  0.27569    
percent_some_college              3.47e-01   2.65e-01    1.31  0.19154    
percent_all_college               3.09e-01   2.65e-01    1.16  0.24449    
Civilian_labor_force_2015        -4.64e-06   6.86e-07   -6.77  1.3e-11 ***
Unemployment_rate_2015            4.29e-02   1.29e-02    3.32  0.00090 ***
pop                               2.17e-06   3.39e-07    6.41  1.5e-10 ***
white_pop                        -1.54e+01   1.71e+00   -9.01  < 2e-16 ***
black_pop                        -1.32e+01   1.72e+00   -7.70  1.4e-14 ***
asian_pop                        -1.59e+01   1.91e+00   -8.33  < 2e-16 ***
indian_pop                       -1.73e+01   1.86e+00   -9.29  < 2e-16 ***
pacific_pop                      -4.12e+01   5.54e+00   -7.43  1.1e-13 ***
hisp_pop                         -1.96e+00   1.95e-01  -10.08  < 2e-16 ***
Poverty_all_population            2.47e-02   1.18e-02    2.09  0.03675 *  
minor_poverty                    -4.06e-03   1.87e-02   -0.22  0.82855    
percent_school_age_children_poverty  3.45e-02   1.76e-02    1.96  0.04981 *  
Household_income                  1.55e-05   2.35e-06    6.58  4.8e-11 ***
avg_wkly_wage                    -4.00e-04   1.10e-04   -3.64  0.00027 ***
labor_force_participation         7.50e+00   5.17e-01   14.50  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.4 on 53306 degrees of freedom
Multiple R-squared:  0.144,	Adjusted R-squared:  0.143 
F-statistic:  191 on 47 and 53306 DF,  p-value: <2e-16
```

Figure 15: OLS Coefficient Summary