Decision 520Q

Team 10: Yaqiong (Juno) Cao, Zhaoji Li, Malika Mohan, Dexter Nguyen, Sleiman Serhan

Student-Crime and Disruption-Related Incident Prediction in the US Public Schools

November 16, 2020

Contents

Business Understanding

Our findings will encompass what may be contributing to higher rates of student-crime and disruption-related incidents at public schools across the US. This information will be beneficial to schools as minimizing these incidents would lead to a better learning environment, less school spending on disciplinary actions, and a potential better school reputation that may invite a higher funding rate. It may also help guide which safety measures are worth investing funds to lower student crimes and offenses. Additionally, they will allow schools to understand the relationships driving student crime/offenses at schools so those schools can best be prepared. The factors explored to do this will be the effects of different types of safety measures schools implement, features of the school, and features of the attending students.

Data Understanding

The dataset came from a survey conducted by the US Department of Education in 2015-2016 and is on crime & safety at a sampled 2,093 public schools across the US. It is described as a "cross-sectional survey of the nation's public schools designed to provide estimates of school crime, discipline, disorder, programs, and policies." The dataset provides information on the school's features (urbanicity, grade levels, size, etc.), safety measures they have in place, and incidents over the past year.

Data Preparation & Exploration

Data Cleaning

After downloading the Department of Education website dataset, we first converted the SAS to CSV data. We then spent time understanding the data and corresponding key of the survey questions schools were asked and evaluating the initial 437 variables. Utilizing domain knowledge, online research, and our business question, we dropped irrelevant/useless columns if they weren't related to our focus area, a repeat, and had meaningless values, leading us to 104 columns. We then renamed the columns to informative titles, re-coded the 2's to 0's for the 'No' dummy variables (so now 0=No and 1=Yes), and re-coded the nulls and skips to all be standardized as "NA." We also created a key for our variables that included the original variable name, our informative title name, the data type, and the question/answers of that variable in the schools' survey.

Next, we counted all the missing values for each variable and built plots to show the missing values by visualization and dropped columns, with over 30% of missing values (Figure 1). Then, we converted the class of numerical variables to categorical variables based on the key survey responses. We grouped columns regarding school policies and programs and feature engineered two variables after researching school reporting and running tests between the two to create our dependent variable of Total Incidents.

Data Exploration and Transformation

Looking at the distribution of our dependent variable Total Incident, we concluded that our data is not normally distributed, but it is slightly skewed (Figure 2). After running a QQ plot, we needed to transform our data to be normally distributed (Figure 2). The new QQ plot demonstrates that our log(Total Incidents) values are much more normally distributed,

allowing us to move forward with our analysis (Figure 2). We also ran BoxCox and had a value very close to 0.

Moving to EDA to understand the characteristics and interactions of different variables of interest, we first used histograms to get an idea of the frequency of School Security measures implemented, counts of average School Procedures/Rules, and the distribution of the numbers of School Prevention Programs available.

Regarding the students' features, we found that the proxies of low academic achievement, including Percent Below 15 Percentile on Standardized Tests, Percent Likely to Attend College, Percent Academic Achievement, was most strongly correlated with Total Incidents (positively for the first and negatively for the latter two). It is reasonable since we may expect lower academic achievement to correspond with higher crime due to less care for school and more time not spent studying/being immersed in their education. After running a model with these three proxies, we saw that Percent Below 15 Percentile (PctBelow15) had the most statistically significant relationship with Total Incidents, so we used that proxy to explore interaction effects. (Figure 3) After running correlation tests on potential interaction variables of the school/student features, we identified Urbanicity, PctWhiteStudents, SchoolAreaCrime, and StudentAreaCrime as having stronger correlations motivating us to investigate them further for interactions. After doing so, we confirmed that we saw high levels of interaction between Student Area Crime, which is crime rates where a student lives, and low academic achievement - which is then correlated with higher Total Incidents at the school itself (Figure 16). It provides us a better understanding of the potential reasoning behind a student's low academic achievement and why that may be translating into them committing more offenses.

Next, to see the relationship between the types of incidents and total incidents, we ran a correlation analysis of a subset of these types of incidents variables against our dependent variable, Total Incidents. We ended up the top 4 variables of interest, which had the highest correlation with Total Incidents: Arrests, VerbalAbuse, GangActivities, and StudentSexualHarassment. Next, we plotted the marginal distributions of the key categorical variables of incidences' types and displayed their relationship with total incidents. Not surprisingly, we found clear positive correlations between Arrests/ GangActivities/ VerbalAbuse/StudentSexualHarassment and Total Incidents (Figure 11-14).

When we look at the relationship between ClassroomChangeAvg - Pct Daily Attendance - PctLimitedEnglish and log(TotalIncidents), accounting for the interaction of School Size, School Size 1,2, and 3 behave similarly. School Size 4, however, is almost always flat or slopes on the opposite side of the other 3 Sizes. (Figure 7 and 8) It means that school size dominates the effect of student features when it comes to incidents reported. Moreover, Figure 9 and 10 shows that a higher percentage of daily attendance leads to a lower amount of arrests on average, but for Urbanicity 4 (Rural Areas), the rate of students' average daily attendance hardly changes the number of arrests, suggesting that schools in rural areas are more prone to Incidents.

To make the dummy variables more comfortable to understand and visualize their relationships with Incidents better, we divided all dummy variables that were indicators of different school safety measures into five categories. We conducted 2-D analysis of these variable groupings against our target variable of interest before analyzing each group against the target variable. We created five new columns to sum up each category:

1. School Training: the total number of School training for each school
2. Law Enforcement: the total number of Law Enforcement for each school

3. Security: the total number of Security for each school
4. Rules: the total number of the Rules for each school
5. Programs Prevention: the total number of Programs Prevention for each school

Looking at our variables' distribution, we concluded that our data is normally distributed (Figure 4-6). The distribution of the School Security feature was normal and symmetric, averaging 7.5 security measures taken per school. For School Procedures/Rules, the average count was five, and an average of 9 School Prevention Programs at each school. (Figure 15)

Correlation Analysis

To see which variables are likely to affect the total incidents, we ran a correlation analysis of all selected independent variables against our dependent variable, Total Incidents. Before doing this, we converted all the categorical variables to dummy variables to allow for numerical calculation of correlation. Once this was completed, we chose to keep the top 10 variables of interest, which had the highest correlation with total incidents (Figure 16).

Last, we considered if the collinearity problem existed in our data. Calculating different correlations among our independent variables and using VIF verification, we discovered a very high correlation (0.8163) between two variables: Gender Identity Harassment and Sexual Orientation Harassment. This, in addition to the output from our previous analysis, was the impetus to remove Gender Identity Harassment from our regression models. We will take advantage of such regularization techniques as LASSO and Elastic Nets in the next modeling part to deal with other highly correlated variables.

Modeling

Based on our EDA and correlation analysis, we used four potential models.

***Model 1*:** From our exploratory data analysis and correlation analysis, we know that Total Incidents is highly correlated with a number of variables (our "Top 10"). We employed multi-linear regression to build an optimal prediction model for the total number of incidents in public schools in the US. We call this "Model 1", which included our "Top 10" explanatory variables (variables highly correlated with Total Incidents). Using K-Fold Cross Validation, we have Model 1 summary as below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-5.0437 -0.6480  0.1092  0.7467  3.9219

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.08441    0.07942  38.836  < 2e-16 ***
SchoolSize_4               0.82214    0.06631  12.398  < 2e-16 ***
GangActivities_4           0.53259    0.07388   7.209 7.87e-13 ***
GangActivities_3           0.93508    0.20172   4.636 3.78e-06 ***
GradeLevel_3               0.15930    0.06457   2.467 0.013709 *
GradeLevel_1              -1.06650    0.06708 -15.898  < 2e-16 ***
StudentAreaCrime_3        -0.22458    0.06763  -3.321 0.000914 ***
SchoolAreaCrime_3         -0.35972    0.07600  -4.733 2.36e-06 ***
GenderIdentityHarassment_4 0.32131    0.05772   5.566 2.94e-08 ***
PctWhiteStudents_4         0.12393    0.05904   2.099 0.035926 *
VerbalAbuse_1              0.81220    0.22456   3.617 0.000305 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.143 on 2081 degrees of freedom
Multiple R-squared:  0.4215,    Adjusted R-squared:  0.4187
F-statistic: 151.6 on 10 and 2081 DF,  p-value: < 2.2e-16
```

In Model 1, all identified variables are highly correlated with our target variable (Total Incidents) and show statistical significance. All the variables have a positive relationship with Total Incidents except Grade Level_1, StudentAreaCrime_3, and SchoolAreaCrime_3. In the case of Grade Level, it suggests that incidents are only popular among high schools (level 3) instead of primary schools (level 1). Meanwhile, a negative relationship with Total Incidents implies fewer incidents in areas (where students live and school located) with a lower level of crime.

Model 2: Next, using the LASSO regression method, we came up with the second model ("Model 2") that performs both variable selection and regularization. This resulted in a subset of

predictors (our "Top 7") that minimizes prediction error for a quantitative response variable - Total Incidents. This subset includes seven variables: GradeLevel_1, StudentAreaCrime_3, SchoolAreaCrime_3, SchoolSize_4, VerbalAbuse_1, GangActivities_3, GangActivities_4. Applying K-Fold Cross Validation again, we got Model 2 summary as below:

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.8720 -0.6758  0.1223  0.7419  3.9211

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.31489    0.06451  51.383  < 2e-16 ***
GradeLevel_1      -1.17221    0.06306 -18.588  < 2e-16 ***
StudentAreaCrime_3 -0.26280    0.06650  -3.952 8.01e-05 ***
SchoolAreaCrime_3  -0.39275    0.07570  -5.188 2.33e-07 ***
SchoolSize_4        0.94870    0.06027  15.740  < 2e-16 ***
VerbalAbuse_1       0.83264    0.22655   3.675 0.000244 ***
GangActivities_3    1.06393    0.20202   5.267 1.53e-07 ***
GangActivities_4    0.60843    0.07289   8.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.153 on 2084 degrees of freedom
Multiple R-squared:  0.4099,    Adjusted R-squared:  0.4079
F-statistic: 206.8 on 7 and 2084 DF,  p-value: < 2.2e-16
```

Using LASSO means fewer variables identified in Model 2. All these seven are highly correlated with our target variable (Total Incidents) and show highly statistical significance at 99%. Like Model 1, all these variables have a positive relationship with Total Incidents except Grade Level_1, StudentAreaCrime_3, and SchoolAreaCrime_3.

Model 3: Besides using Lasso to reduce overfitting in our linear model, we also considered Elastic Net, which combines feature elimination from Lasso and the feature coefficient reduction from the Ridge (another regularization technique) to improve our model's predictions. Elastic Net helps us to have new insights into the StudentBullying_2 variable. K-Fold Cross Validation resulted in the model's summary, as shown below, with all the independent variables significant.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.6872 -0.6552  0.1209  0.7628  3.4508

Coefficients:
                   Estimate Std. Error t value             Pr(>|t|)
(Intercept)         3.12692    0.07502  41.683 < 0.0000000000000002 ***
StudentBullying_2   0.44408    0.07587   5.853      0.00000000559355 ***
VerbalAbuse_2       0.53664    0.11623   4.617      0.00000412759351 ***
VerbalAbuse_1       0.90987    0.22378   4.066      0.00004961551536 ***
GangActivities_3    0.92375    0.20052   4.607      0.00000433524520 ***
GangActivities_4    0.51515    0.07355   7.004      0.00000000000334 ***
StudentAreaCrime_3 -0.21649    0.06731  -3.216               0.00132 **
SchoolAreaCrime_3  -0.32549    0.07599  -4.283      0.00001924861699 ***
GradeLevel_1       -1.15237    0.06268 -18.384 < 0.0000000000000002 ***
SchoolSize_4        0.94336    0.05948  15.859 < 0.0000000000000002 ***
PctWhiteStudents_4  0.10180    0.05852   1.740               0.08208 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.137 on 2081 degrees of freedom
Multiple R-squared:  0.4277,    Adjusted R-squared:  0.425
F-statistic: 155.5 on 10 and 2081 DF,  p-value: < 0.00000000000000022
```

Model 4: Last, we decided to run Random Forest as a machine learning regression tree algorithm used in the modeling process. This helps to create a random sample of multiple regression decision trees and merges them to obtain a more stable and accurate prediction through cross-validation. We call this "Model 4", with its summary as below:

```
Call:
 randomForest(formula = log(TotalIncidents + 1) ~ ., data = training,      mtry = 3,
importance = TRUE, na.action = na.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 1.130716
                    % Var explained: 50.57
```

Diving deep into variable selection, we have the top 10 predictors most important to the model. It is done by using MDI (Gini Importance or Mean Decrease in Impurity) that calculates each feature's importance as the sum over the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits. In comparison with Model 1, Model 2, and Model 3, we have additional insights into such variables as FTSecurityCount,

PctAcademicAchievement, PctLikelyCollege, PctAverageDailyAttendance, and PctSpecialEd (Figure 17). This finding also matches with our prior analysis from the EDA part.

Evaluation

After running our four models, we used three metrics: R-squared, RMSE, and MAE, to evaluate our model prediction performance. As we expected from Figure 18, Model 4 is the best in terms of all three metrics, with R-Squared: 50.57%, RMSE: 1.016, and MAE: 0.7879. Model 1, Model 2, and Model 3, whose predictors selected from our correlation analysis and regularization techniques like Elastic Net and LASSO, meanwhile, don't record much difference in terms of these performance metrics. (Figure 18 and 19)

It is reasonable that Random Forest in Model 4 gives us superior "predictions". However, from a perspective of "marginal impact" interpretation, Model 1, 2, and 3 may be the winners even though their measured performance is behind. In the context of our business question, focusing on predicting the total number of incidents, Model 4 will be the best choice.

Implications, limitations, and conclusion

The chosen model will help serve the US government and public schools as they will wield it to predict schools with higher levels of crime, and allocate funds and enact policies accordingly. Our key findings showed that there were recurrent School Features that appeared in each of our four models. These include School Size 4 and Grade Level 1. This first variable means that most schools with 1000+ students are the ones most prone to Incidents. The negative correlation of Grade Level 1 (Primary Schools) with Total Incidence implies that Primary Schools are not exposed to high levels of Incidents. Next, Student Area Crime 3 and School Area Crime 3, with negative correlations with Total Incidents, implies that schools that are and have

students from decent neighborhoods are less likely to be prone to incidents happening in their respective institutes. Furthermore, the first three models show a strong relationship with trouble being caused mainly by Gang Activities and Verbal Abuse. Consequently, we now have a clear picture of what factors are least/most likely to lead to reported incidents.

First, we can confidently eliminate Primary Schools and consider only Middle and High Schools for our final recommendations. The higher the School Size, the more likely incidents will happen due to the unpredictable nature of managing large groups. Second, we can also safely eliminate from our analysis the areas/schools/students that represent "decent" neighborhoods since these are areas with low crime rates and a high average salary per household. Finally, we can make sure to deploy significant funding into training, workshops, and prevention programs that limit the presence of Gang-related activities and Verbal Abuse. Thus, a typical candidate for federal funding and support would be a large 1000+ students High School that serves a population of lower-income families, where the presence of Gang Activities is high, and the Percentage of White Students is low.

These findings were validated across many ways, including correlation analysis, two-sample t-tests, and regression modeling. Our selected Model 4, which uses the random forest to generate the most influential variables related to Total Incidents, confirms our hypothesis that a large part of Total Incidents is often explained by the students' characteristics and long-term career goals and ambition. One of the limitations of this model is that as previously shown in our EDA, is that sometimes even if the students come from good backgrounds, the random nature of large schools makes it highly likely that Incidents of some kind will happen.

Our findings also help guide what schools should enact. Largely populated institutions from lower-middle class neighborhoods with high crime rates are the most statistically

significant measure areas to pay attention to the impact of student crime levels, particularly at the Middle/High School level. However, the most accurate student crime levels indicators will not come from the safety measures in place. Instead, school/student features like a high percentage of students who consider academic achievement essential and are likely to go to college, Middle-High Schools, and Full Time Security Count are the strongest predictors. Schools should be armed with this knowledge to correctly identify suspect characteristics, then target the correct types of frequent incidents before designing and implementing prevention measures and training.

There are some limitations of this dataset, and thus findings that are important to note. Firstly, there may have been reporting bias present. School administrators filled out surveys and may have underreported the number of incidents/disciplinary actions if they were embarrassed or concerned about how often these things happened. Additionally, some of the data collected may have been subjective such as 'number of hate crimes reported' as different schools may classify hate crimes differently. Or data might be predisposed to have certain prejudices/biases that may factor into their classification of how seriously they take a crime. Lastly, these models need to be analyzed with more subject matter expertise in the education domain to implement our suggestions effectively as different schools may interpret them and enact them differently. For example, each school can have different quality levels of training programs, and this variation in quality could significantly affect the impact of their use. Another issue could be that urbanicity doesn't appear in any model. Meanwhile, from our EDA, Rural areas are more likely to have Incidents than other areas. Thus, we can conclude that models cannot paint the whole picture for some unknown reasons and should be combined with one with domain knowledge and the insights from our EDA.

Appendix: Referenced Figures

Figure 1

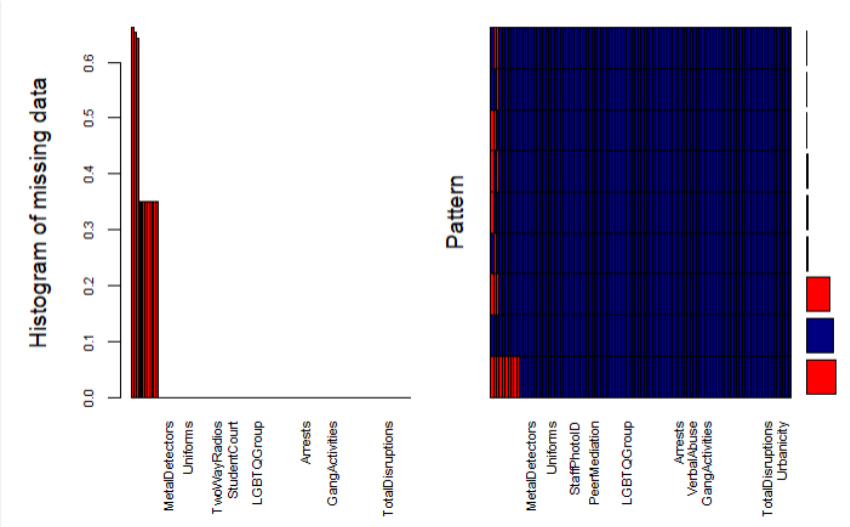
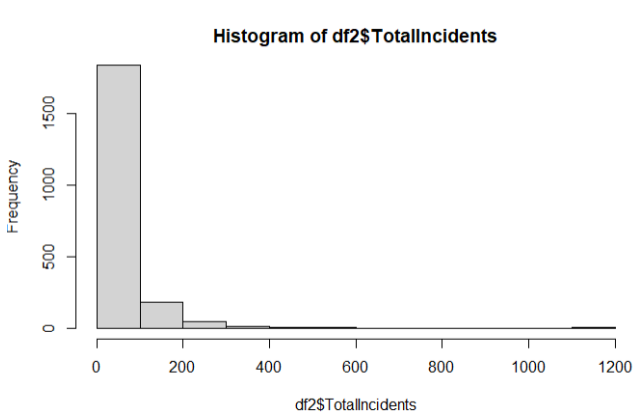
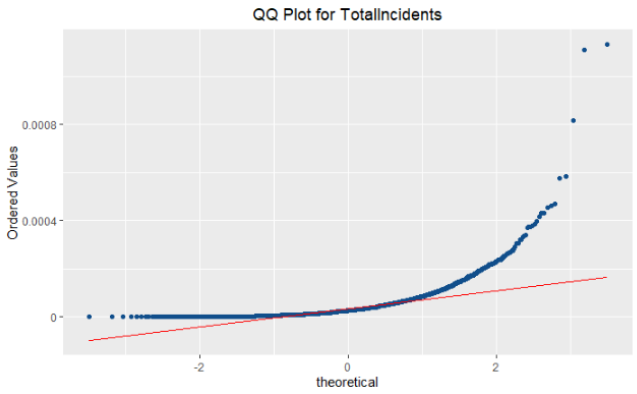
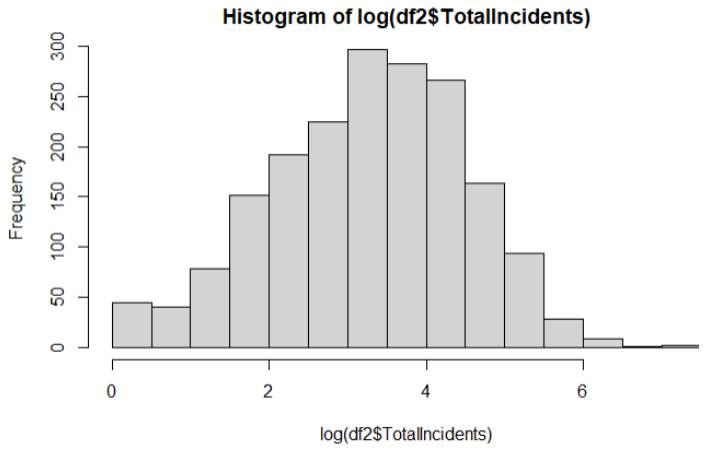


Figure 2







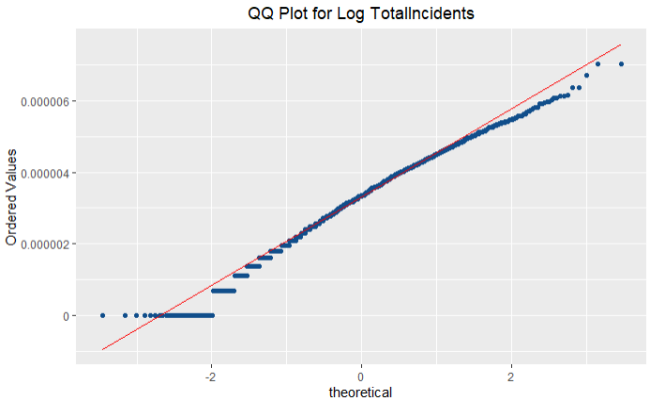


Figure 3

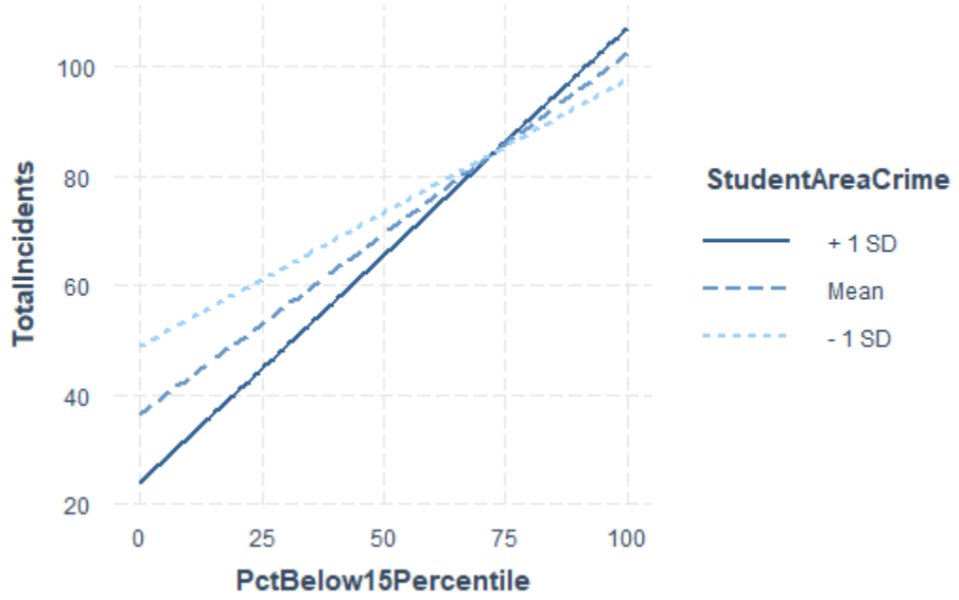


Figure 4

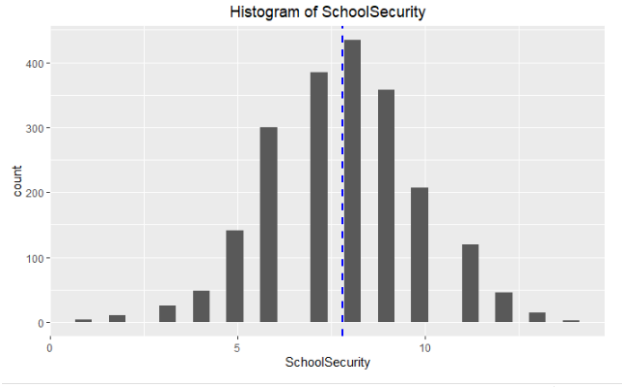


Figure 5

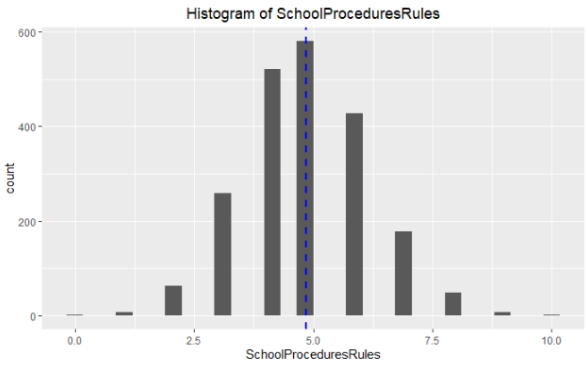


Figure 6

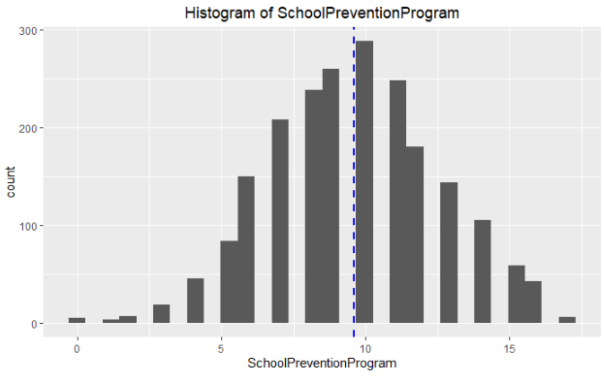


Figure 7

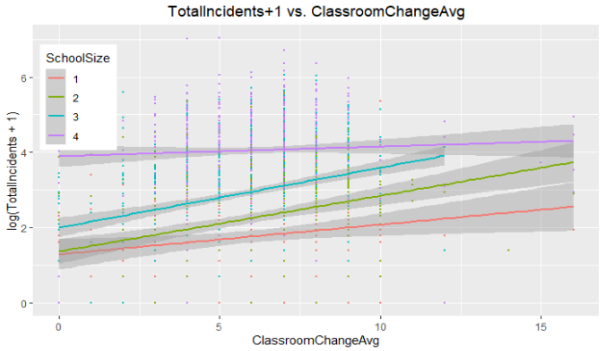


Figure 8

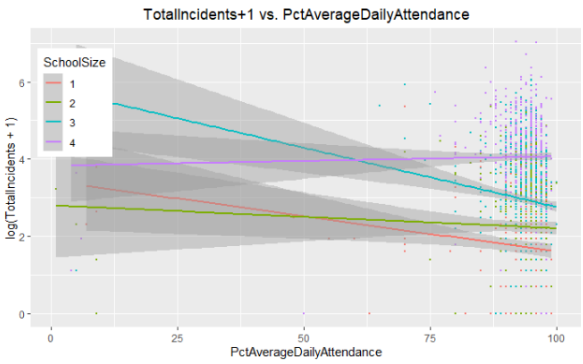


Figure 9

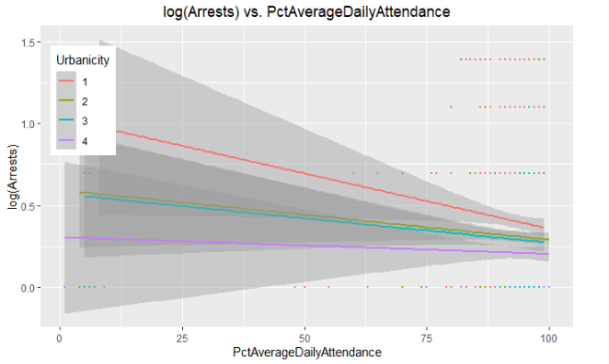


Figure 10

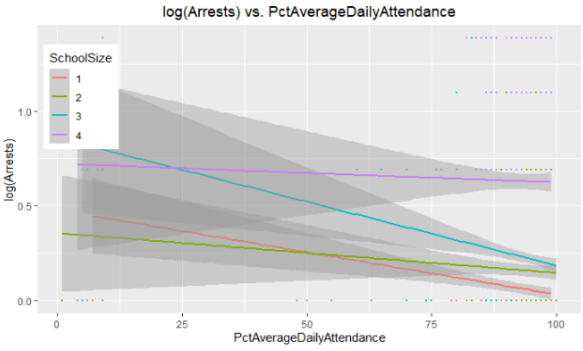


Figure 11

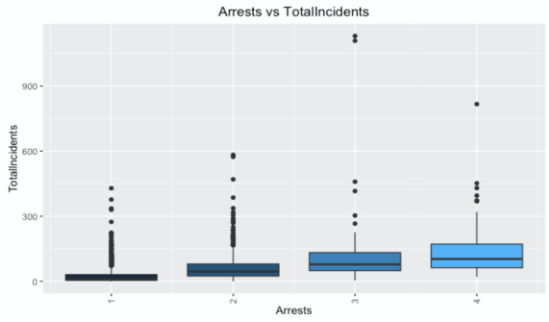


Figure 12

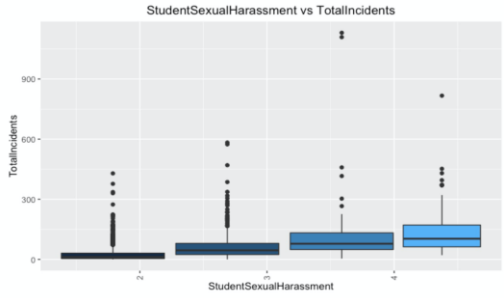


Figure 13

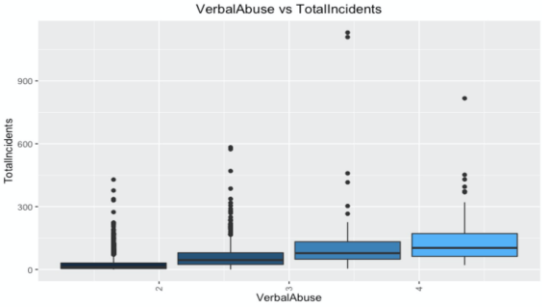


Figure 14

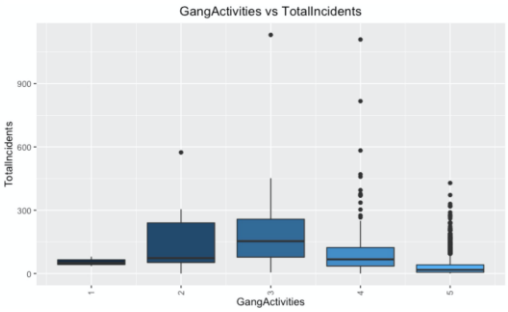


Figure 15

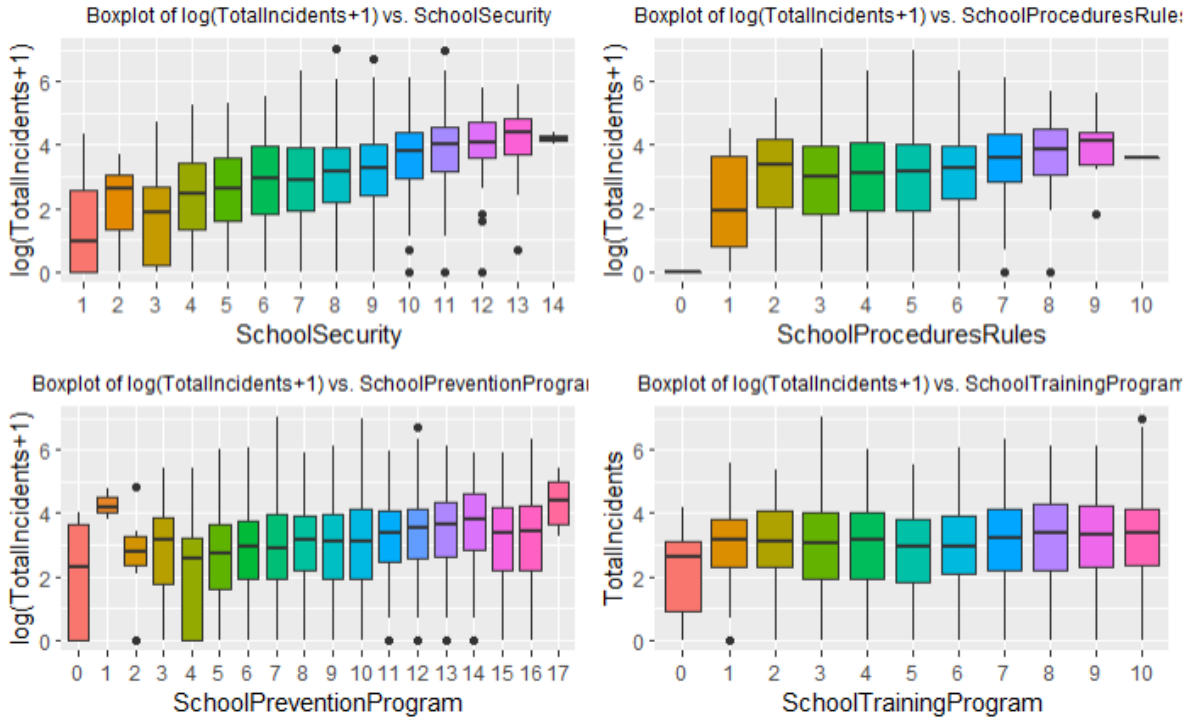


Figure 16

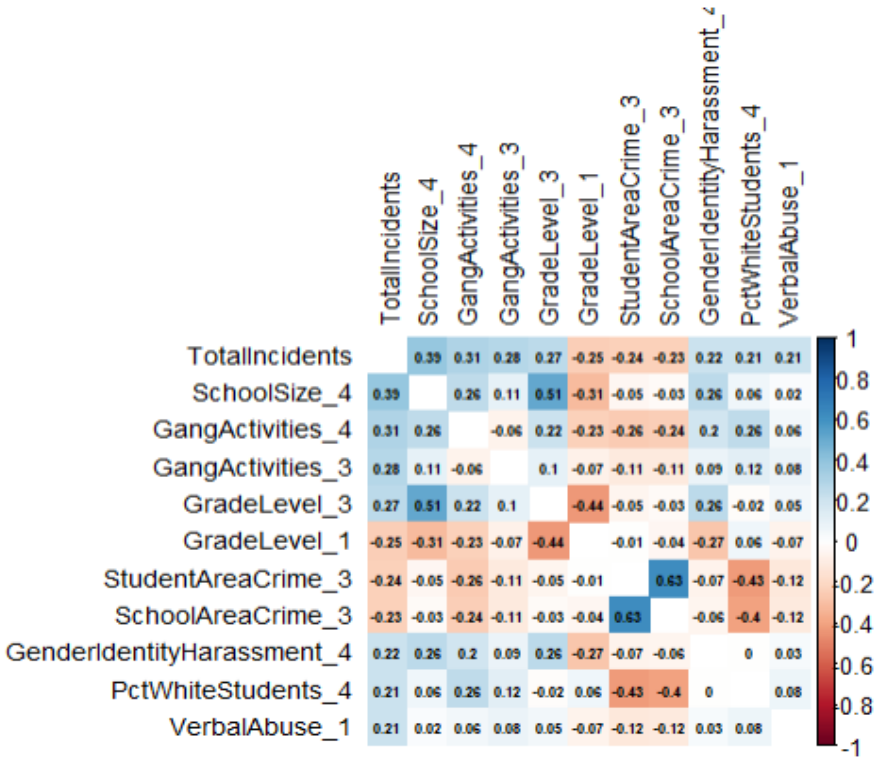


Figure 17

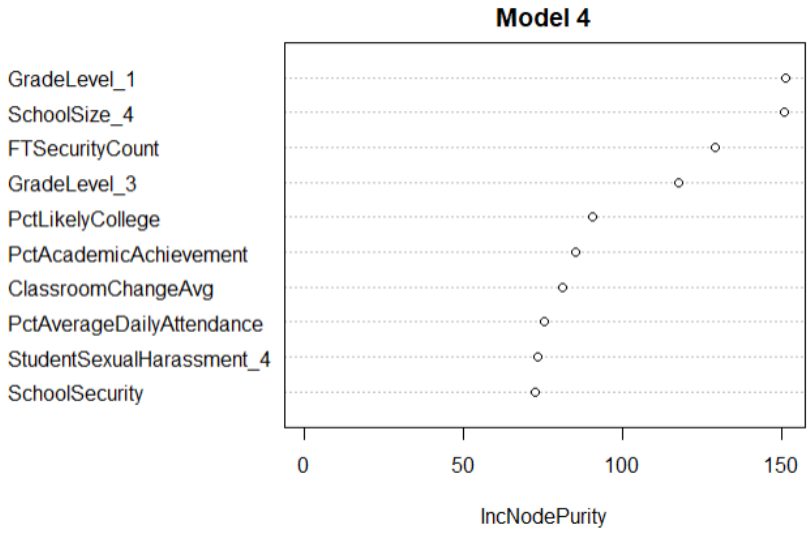
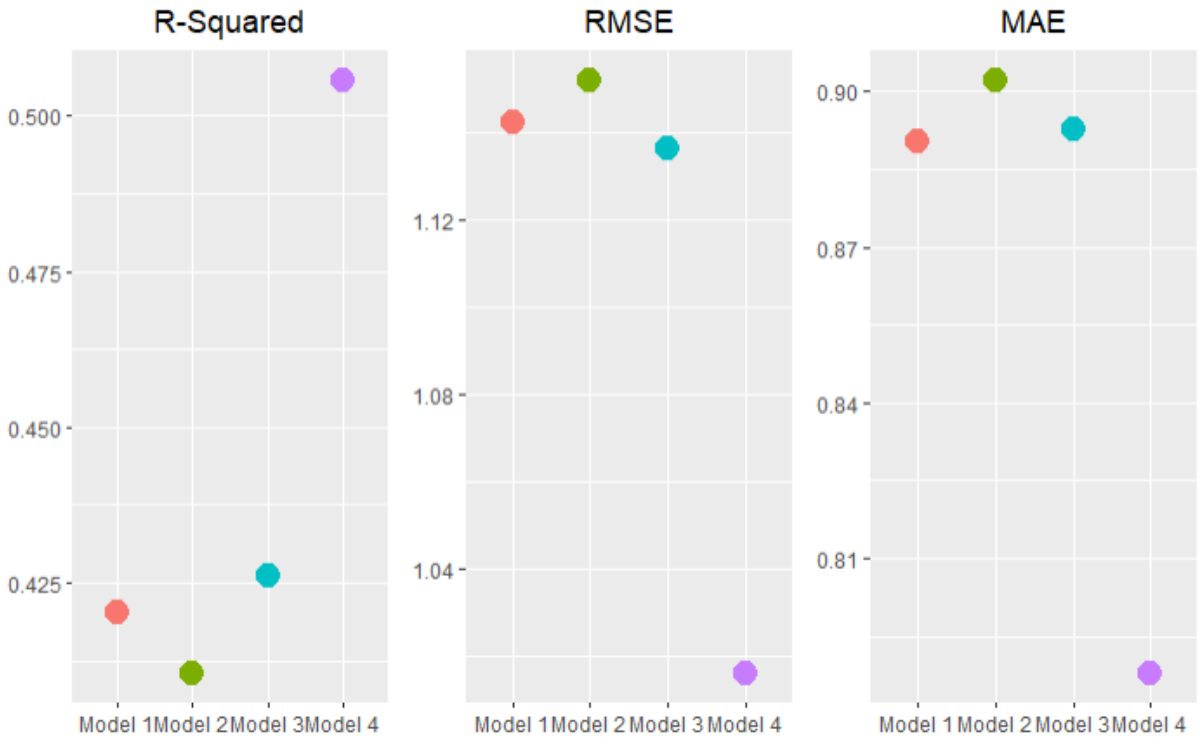


Figure 18

Variable Category / --- ----- Model ---------	**Student Features**	**School Features**	**Types of Incidents/ Surveys**
Model 1 (10 Highest Correlations)	• Student Area Crime 3 • Pct White Students 4	• School Size 4 • Grade Level 1 • Grade Level 3 • School Area Crime 3	• Gang Activities 3 • Gang Activities 4 • Gender Identity Harassment 4 • Verbal Abuse
Model 2 (LASSO)	• Student Area Crime 3	• Grade Level 1 • School Size 4 • School Area Crime 3	• Gang Activities 3 • Gang Activities 4 • Verbal Abuse 1
Model 3 (Elastic Net)	• Student Area Crime 3 • Pct White Students 4	• School Size 4 • Grade Level 1 • School Area Crime 3	• Gang Activities 3 • Gang Activities 4 • Verbal Abuse 1 • Verbal Abuse 2 • Student Bullying_2
Model 4 (Random Forest)	• Classroom Change Avg • Pct Academic Achievement • Pct Likely College • Pct Daily Avg Attendance • Pct Special Education	• GradeLevel 1 • School Size 4 • FT Security Count • Grade Level 3	• Student Sexual Harassment 4

Figure 19


R-Squared
0.500
0.475
0.450
0.425
Model 1
Model 2
Model 3
Model 4
RMSE
1.12
1.08
1.04
Model 1
Model 2
Model 3
Model 4
MAE
0.90
0.87
0.84
0.81
Model 1
Model 2
Model 3
Model 4