School Violence in the US (2010-2018)

Investigation by Tyler Schmalz

Introduction

According to the Academy for Critical Incident Analysis, between 2000 and 2010, there were 57 incidents of school violence worldwide that had two or more victims. Twenty eight (almost half!) of those occurred in the United States alone. Over the past few decades, students and teachers have expressed a growing concern about safety at schools, specifically related to these incidents. There are a variety of opinions on how best to reduce or protect against these events: increase in security, frequent drills to prepare, mental health outreach, etc. However, for each suggestion there is always the question of how feasible it is in terms of resources (time and money) to implement nationwide.

This question lead to my analysis of school shooting incidents in the United States from 2010-2018. It is my purpose through this study to determine if there are early warning signs or common characteristics of areas with school violence in the United States. More specifically, I am interested in the following questions:

1. Which variables have the largest positive or negative impact on the likelihood of school violence?
2. Is school violence more likely to be clustered in similar areas/ times, or more uniformly distributed?
3. What are the (if any) common characteristics of the schools or counties in which you find the majority of recent school violence?
4. What relationship (if any) is there between a county's overall adult/teen health and the likelihood of a school shooting occurring?

Motivation

I believe that this analysis is of interest to a variety of organizations and clients, regardless of their opinion how to best approach reducing the occurrences of school violence and protecting students. If there is a way to accurately tag counties that are more likely to experience school violence, these organizations could more efficiently focus their resources to have a greater impact. Also, the discovery of features with particularly strong connections to school violence could inform more targeted interventions. It is my hope that my analysis is the beginning of reducing the frequency and severity of incidents of school violence.

Data

My analysis is based on the following sources:

1. The Washington Post's dataset records information on school violence since Columbine in 1998. It is easily downloadable as a .csv file from the link as well as clear documentation on the variables included.
2. County Health Rankings records information and ranks counties in the United States based on various characteristics from 2010 to current. It is also easily downloadable as a .csv file from the link with clear documentation on the variables included.

While the county health rankings had a variety of variables for which information was recorded, I opted to select a smaller sample for my initial investigation: average mentally unhealthy days per week, percent who report binge drinking, teen birth rate, percent uninsured, high school graduation rate, percent of adults with some college education, unemployment rate, percent of children living in poverty, percent of single-parent households, and the primary care physician ratio to the county's population. There were other variables that I was interested in including (such as access to healthy food, mental health support, etc.) but these were not consistently reported throughout the years.

Data Cleaning

The school shooting and the county health datasets were merged together using FIPS county codes which are unique to each county. Then, the following steps were performed to clean and prepare the data for analysis:

1. Dropped information:
    a. Columns: age_shooter2, gender_shooter2, race_ethnicity_shooter2, shooter_relationship2, shooter_deceased2, deceased_notes2, deceased_notes1, weapon_source
        i. The columns about a second shooter had only 2 entries, leaving 105 empty values in each.
        ii. Didn't feel that deceased_notes1 (how the shooter died) would add much value to later analysis
        iii. Weapon_source (where the student got the weapon) was mostly missing and I don't believe it will add a lot to later analysis
    b. Row for the school shooting at Vereen School on November 5, 2015
        i. This row was responsible for the only missing value in many columns (school demographics, shooter information, and time).
2. Missing values:
    a. 22 missing values for the age of the shooter
        i. Replaced with the median age of shooters
        ii. Chose median because there is a wide range of ages (from 6 to 53)

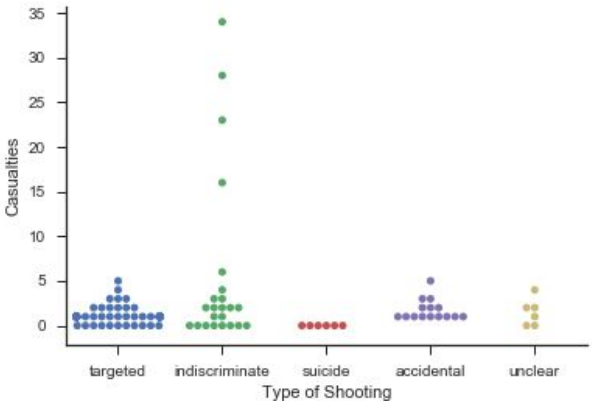
b. Various numbers of missing values in columns state, school_type, shooting_type, gender_shooter1, race_ethnicity_shooter1, ulocale, day_of_week, and city
    i. Replaced all missing entries with string ‘unknown’ in preparation for converting to categorical variables.
c. 2 missing times of the incidents
    i. For Stellar Leadership Academy, I replaced the missing value with 1:30 PM. I was able to find this rough approximation in a local news article written about the event.
    ii. For Redland Middle School, I replaced the missing value with 11:00 AM. There was no time reported in any of the news articles, but they all mentioned that it happened during a student’s class. So I chose the middle of the typical school day as the time.
d. Various number of missing values for racial/ethnic demographics of the school
    i. Replaced missing counts for Rebound High School and Success Academy based on online reports.
    ii. Assuming all remaining missing ethnic counts (hawaiian native and two or more ethnicities) are actually 0
e. Missing values in staffing and lunch
    i. Replaced all missing values in staffing and lunch with 999 and 9999 respectively
        1. These represent unknown values while maintaining the columns as numeric dtypes

3. Correcting dtype of DataFrame columns
    a. Categorical
        i. Converted state, school_type, shooting_type, gender_shooter1, race_ethnicity_shooter1, shooter_relationship1, weapon, ulocale, day_of_week, and city
            1. shooter_relationship1 - There were 20 unique responses in this column that I re-grouped into student, former student, student relation, not a student, and unknown
            2. weapon - There were 20+ unique responses in this column as well so I regrouped them into generic type of gun (revolver, handgun (semi-automatic), shotgun, rifle, unknown
            3. shooting_type - Consolidated responses into fewer categories (targeted, accidental, indiscriminate, suicide, and unclear)
    b. Datetime
        i. Combined date and time columns into a new column labeled date_time. Then converted date_time column to a datetime type.
    c. Numeric
        i. Replaced kindergarten(KG) and pre-kindergarten(PK) in low_grade column to numeric values 0.5 and 0 respectively.
        ii. Then converted both low_grade and high_grade to float type.

Exploratory Data Analysis

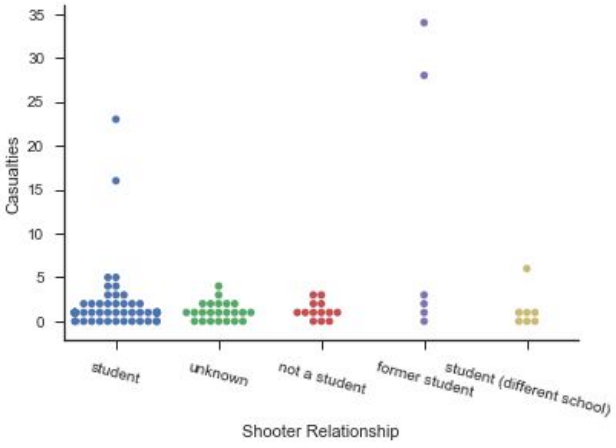
Targeted vs. Indiscriminate

I began my analysis by investigating the breakdown of incidents into their type of school shooting: targeted, indiscriminate, suicide, accidental, or unclear/unknown. In the figure to the right, we can see that the targeted shooting types comprised the majority of incidents (with each dot representing a single incident). While targeted shooting incidents were the most common, they had relatively low casualty rates per incident. Indiscriminate shooting types, while less common, had the most devastating impact with 4 of the incidents having more than 15 casualties. In fact, the indiscriminate shootings had 129 casualties altogether while the other types combined had 99. While the other types of shootings are equally important in terms of prevention and safety, I chose to focus on targeted and indiscriminate shooting types as they contained the vast majority of incidents as well as casualties.



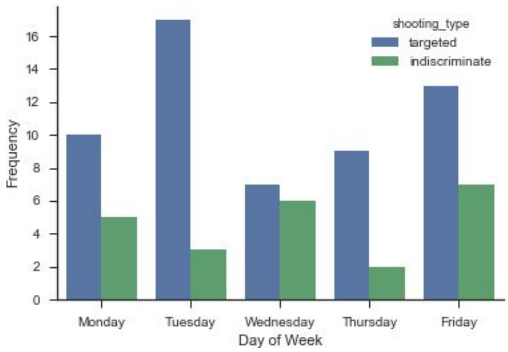
Shooter's Relationship to School

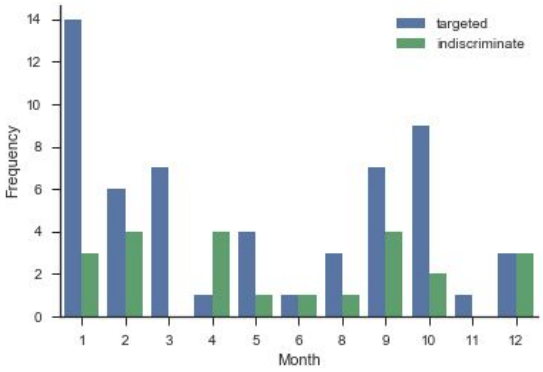
Next, I was interested in the relationship of the shooter to the school. As seen in the figure to the left, the majority of school shootings are perpetrated by a student, most of those by a current student of that school. And the incidents with the highest casualty rates are committed by a current or former student of that school. While any incident of school shooting is important and should be investigated thoroughly, it appears that the most common and devastating of these are perpetrated by students targeting an individual or individuals, or indiscriminately attacking individuals. If we look into these incidents more closely, what might we find?



Day and Month of Incident

There appears to be a difference between the two shooting types in relation to the day of the week that the shooting occurred on. Indiscriminate shootings appeared to occur independently of the day of the week, while the targeted shootings seemed to mostly favor Tuesday. However, after calculating a χ^2 goodness of fit statistic for both, there wasn't any significant statistical evidence to suggest that they were not evenly distributed. (Targeted: p-value = 0.2461, Indiscriminate: p-value = 0.4425)





There were, however, more substantial results when looking at the distribution of months for each type. For the most part, the results were as one would expect. There is a substantial dip during the summer months (June through August) while most schools are traditionally not in session. It also appears that indiscriminate shootings (green bars) are fairly evenly distributed throughout the remaining months. However, it appears that targeted shootings are most likely to occur in January. In fact, almost 18% of all school shooting incidents during this time period occurred were targeted shootings in January!

There is statistically significant evidence through a χ^2 goodness of fit test to support the hypothesis that the months are not evenly distributed for targeted shooting types (Targeted: p-value = 0.0025, Indiscriminate: p-value = 0.8494). Upon further investigation, through a hypothesis test for proportions, there is also statistically significant evidence to support that January has an unusually higher proportion (z = 2.57, p-value = 0.0102) than you would expect the incidents were occurring independently of the month. Surprisingly, there were also two months that were substantially lower than expected: November (z = -4.824, p-value = 0.000001) and April (z = -4.824, p-value = 0.000001).

Clearly, there is something important about these months or the months leading up to it. Perhaps there is some relation to the traditionally big family holidays adding stress to a student's life, or the stress of grades being released over winter break?

County Health Statistics

Next, I searched for statistically significant differences between the various health characteristics per county. I continued on this track of comparing and contrasting targeted vs. indiscriminate shooting types to find that there may be some significant differences between the counties in which these occurred. Most notably, the counties with targeted school shootings had substantially higher teen birth rates, uninsured individuals, and the number of children in poverty.

Lastly, I compared counties with higher casualty rates (greater than 1) and counties with lower casualty rates (1 or lower) per each incident to find that there is little evidence to suggest they are different with respect to the variables I included. However, there was one substantial difference: self-reported rates of excessive drinking. I found it particularly interesting that there didn't appear to be a difference in the other variables that many people would attribute to an increase in drinking: mentally unhealthy days, unemployment, etc. What about heavy/binge drinking in parents could be related to a higher casualty rate in incidents of school shootings?

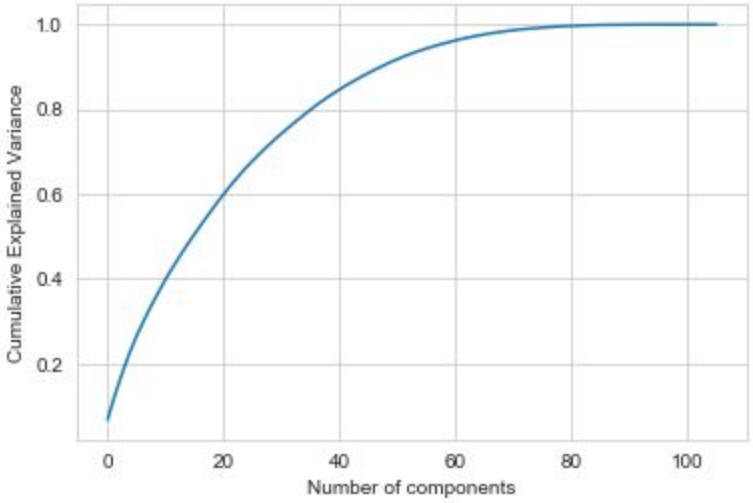
To summarize, it appears that there is a strong relationship between the type of school shooting, the frequency in which it occurs, and the casualty rate, as discussed in my earlier sections. It also appears that frequency of school shootings is substantially higher in January than would be expected, and lower in November/April than would be expected. When looking at the county health characteristics, it emerged that there may be a relationship between a county's teen birth rate, percentage of uninsured individuals, and the number of children in poverty to the type of school shooting, especially when differentiating between targeted and indiscriminate incidents.

In-depth Analysis

In earlier exploratory data analysis, it appeared that there was a strong relationship between the type of school shooting, the frequency in which it occurs, and the casualty rate. It also appeared that the frequency of school shootings is substantially higher in January and lower in November/April than would be expected. The earlier exploratory data analysis supports the hypothesis that these features are strongly connected with the severity of the event (as measured in the casualty rate). However, I plan to use supervised and unsupervised machine learning algorithms to further test these hypotheses.

Feature Selection

Because of the large number of categorical variables and categories, the resulting dataset included 100+ features. This lead me to use principal component analysis to determine if a lower number of features could represent the information nearly as reliably. The following plot shows the relationship between the number of components/features included, and the cumulative ratio of variance in the data that could be explained.



From the image above, it appears that 45 components can account for about 90% of the explained variance. Therefore, I decided to use the 45 components from the principal component analysis to predict the casualty rate of an event.

Predicting Severity of Event

My next steps were to separate my PCA components and casualty rate into train and test sets. I decided to use Ridge regression to predict the casualty rate based on the PCA components as explanatory variables. I started by using a grid search cross-validation to determine the optimum alpha value for the Ridge regression. Once that was determined, I fit the Ridge regression model to the training data, and calculated a score to evaluate the accuracy of the model. The model was returning a coefficient of determination anywhere from 50-60%,

meaning that my model could account for about 50-60% of the variation in the casualty rate. Therefore, my model is moderately accurate at predicting casualty rates.

Identifying strong features

The second aspect of my linear model for predicting casualty rate that I am interested in is which variables seemed to have the largest impact, positive or negative, on the casualty rate. In the table below are the highest correlation coefficients, demonstrating the strength and direction of the relationship between these features and the casualty rate.

Feature	Correlation Coefficient
State: Connecticut	0.5152
Shooting Relationship: Former Student	0.4593
Shooting Type: Indiscriminate	0.3717
Children in Poverty	-0.2002
Teen Birth Rate	-0.2049
Shooting Type: Targeted	-0.2179

The large positive correlation coefficients suggest that schools in Connecticut, former students of the school, and an indiscriminate shooting type typically have a higher casualty rate. Interestingly enough, it also provides evidence that counties with higher rates of children in poverty and teen birth rates typically have lower casualty rates in their school violence. However, there are two caveats to these conclusions:

1. All of the correlations show a moderate to weak relationship
2. These values could be impacted by the two incidents with much higher casualty rates than the other incidents.

To further investigate the relationship between the features and the casualty rate, I continued with a Lasso regression model. Lasso regression helps to identify strong features because it will essentially reduce coefficients of the weaker features to zero. After I fit the Lasso regression to the sample data, I found there were only two features with non-zero coefficients: the number killed and the number injured. Clearly, those would play a strong role in predicting the casualty rate. I thought that maybe the connection between those two features and the casualty rate affected the coefficients of the other variables. I removed them from the sample data, and re-ran the same steps to refit the model and calculate the coefficients. Unfortunately, the Lasso regression model still reduced all of the coefficients to a 0 value, suggesting that none of the features have a particularly strong relationship with the casualty rate.

There is conflicting evidence about the strength of the relationship between some of the features and the casualty rate. While the correlation coefficients support earlier hypotheses that indiscriminate shooting types and former students are strongly related with higher casualty rates, the Lasso regression suggests that none of the features are very closely related to it.

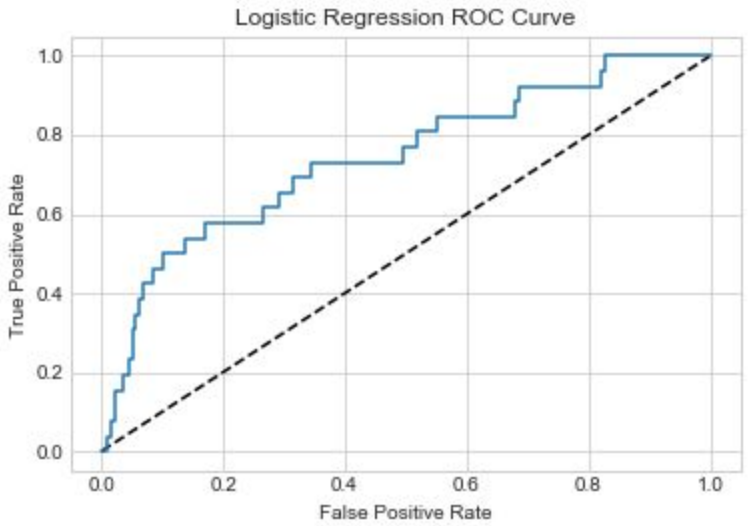
Investigating county relationship to incidents of school violence

My investigation using the characteristics of the individual schools/counties in which school violence occurred lead to inconclusive results, which I believe is a result of having only 100 observations from 2010-2018. To increase the sample size, I zoomed out to focus on the county health characteristics during that time period and fit a model to predict how likely a county was to have an incident of school violence.

In order to do this, the county health dataframe first needed some minor cleaning. I included a column that indicated whether there was at least one incident of school violence in the county during the year. I replaced missing values with the median value for the county over the years 2010-2018. If all of the values were missing, I then used the median value of all counties for that specific year.

In order to prepare the dataframe for a logistic regression to calculate the probability of an incident of school violence occurring, I used the StandardScaler() function from the sklearn package. I then split the dataframe into testing and training sets stratifying to make sure that a similar percentage of both categories fell into the training and testing sets.

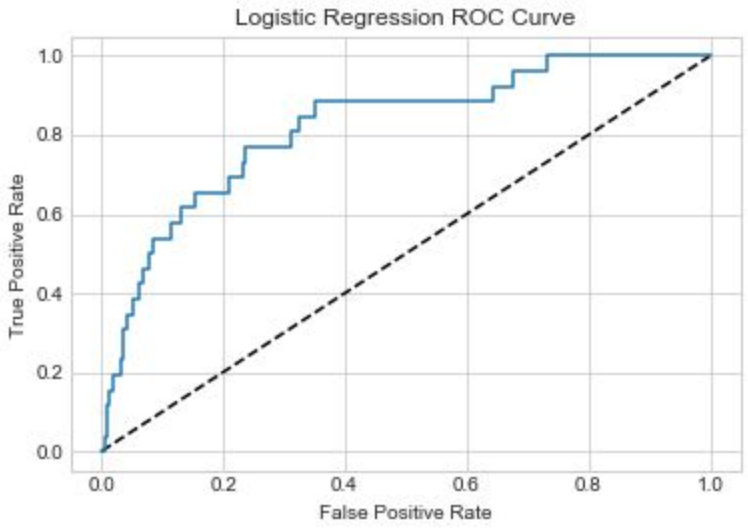
I applied the standard logistic regression to the training data, and calculated the following ROC curve. This ROC curve suggests that my logistic model does fairly better than a truly random model.



However, all of the probabilities were extremely low (8% or lower) and the model was predicting no school violence for every case. While this lead to a high overall accuracy, it wasn't very

helpful. Upon further investigation, the incidents of school violence only comprised about .3% of the overall dataset, which is called the rare event problem. In order to manage the rarity of a school violence event, I used the smote object from the imbalanced-learn object in order to upsample the rare events, so that there was 50-50 split between both events.

I then refit a similar Logistic Regression model to the resampled data. I recreated the ROC curve from earlier, and it seems like this new model performs slightly better with the resampling, as suggested by the increased area under the ROC curve.



However, there is still the issue that that the model is predicting the same event (no school violence) for every situation. While this is leading to a high accuracy again, it's clearly not an effective model. The model is still predicting very low probabilities for each observation with the highest only around 20%.

Conclusion

In this investigation of school violence in the United States from 2010-2018, there were several interesting realizations about its connection to other factors, some particular to the school and others to the county as a whole. For example, the highest casualty rates all belonged to the "indiscriminate" shooting category, while the majority of all events were "targeted" but with smaller casualty rates, suggesting that interventions may be better focused on these types of school shootings. It was also surprising that targeted/indiscriminate school shootings occur much higher during January than one would expect. Possibly, there is a connection between the post-holidays high and a student's mental health.

However, the in-depth analysis using supervised and unsupervised machine learning algorithms was more inconclusive. The Ridge regression model struggled to accurately predict the casualty rate of an incident, and the Lasso regression showed no strong features in relation to the casualty rate as well. The Logistic regression model also struggled to predict the likelihood of an incident of school violence occurring.