2025 US School Shooting Predictions and Solutions*

Leveraging Forecasting to Shape Policy and Protect Communities

Yun Chu

November 28, 2024

School shootings in the United States pose a critical challenge with far-reaching impacts. This paper employs a Random Forest regression model to predict state-level casualties for 2025 using historical data from 1999 to the present. The findings reveal a positively skewed distribution of casualties, with states like Georgia, California, and Pennsylvania projected to face the highest numbers. These insights highlight the need for targeted interventions, including enhanced school safety measures and stricter gun control policies. By providing actionable predictions, this study aims to inform policies that reduce school shooting casualties and protect vulnerable populations.

Table of contents

*Code and data are available at: https://github.com/chuyun2024/School-Shooting-Analysis.

Introduction

School shootings remain a grave concern in the United States, with devastating consequences for students, educators, and communities. Despite the growing public discourse and policy debates surrounding school safety, little progress has been made in effectively predicting and preventing these tragic incidents. Understanding the patterns and predictors of school shootings is essential for designing evidence-based interventions to protect students and reduce casualties. This paper contributes to this critical issue by leveraging data-driven methodologies to analyze and predict school shooting casualties, with the aim of informing policy and prevention efforts.

The estimand in this paper is the number of casualties from school shootings across U.S. states for the year 2025. Using historical data from The Washington Post (The Washington Post 2024b), which provides a comprehensive record of school shootings, a Random Forest regression model was developed to predict casualties based on state and temporal features. The analysis includes summary statistics, spatial visualizations, and predictive modeling, providing actionable insights for policymakers. This study fills a critical gap in understanding the state-level distribution of school shooting casualties and offers data-driven recommendations for targeted interventions.

A key finding of this study is the uneven distribution of casualties across states, with some states, such as Georgia, California, and Pennsylvania, predicted to experience significantly higher casualties in 2025. This highlights the importance of state-specific policies and resource allocation. The analysis also reveals that the distribution of casualties is positively skewed, with most incidents involving relatively few casualties, but a small number of high-casualty events disproportionately contributing to the overall impact. This information underscore the need for targeted strategies to mitigate the effects of high-casualty incidents.

The importance of this study lies in its ability to provide a data-driven foundation for policy and prevention efforts. By identifying high-risk states and understanding the characteristics of high-casualty incidents, this paper aims to guide policymakers in allocating resources effectively and designing interventions that address the underlying factors contributing to school shootings.

The structure of this paper is as follows: Section discusses the data, including its sources, measurements, and key variables; Section outlines the model used for prediction, including its justification and validation; Section presents the results, including summary statistics, spatial visualizations, and predictions for 2025; and Section concludes with a discussion of the findings, limitations, and recommendations for future research and policy.

Data

Overview

The dataset has 416 entries, with each entry representing a unique school shooting incident. Incidents occurring during after-hours events, accidental gun discharges that only injured the individual handling the firearm, and private suicides that did not endanger other children were excluded from consideration. Additionally, shootings at colleges and universities, which involve young adults rather than children, were not included in the analysis (The Washington Post 2024b). These entries cover 50 variables that provide information about the schools, locations, date and time of shooting, shooters details, number of killed and injured, and the relationship of the shooter to school, the weapon type and source.

As the federal government does not consistently track school shootings, this dataset from *The Washington Post* fills a critical gap. It was carefully assembled using information from diverse sources, including news articles, open-source databases, law enforcement reports, and direct inquiries to schools and police departments. Although sources like FBI crime reports and local school incident logs were reviewed, they lack the detail and comprehensive coverage of this dataset. Its unparalleled breadth and depth make it the strongest foundation for predictive modeling and generating actionable insights.

We use the statistical programming language R (R Core Team 2023) to download, clean, analyze and model the US School Shooting Data. The US School Shooting dataset is downloaded

from The Washington Post (The Washington Post 2024a) . The following libraries are utilized in this paper:

- tidyverse (Wickham et al. 2024e)
- dplyr (Wickham et al. 2024a)
- lubridate (Grolemund and Wickham 2024)
- readr (Wickham et al. 2024b)
- stringr (Wickham et al. 2024c)
- arrow (Richardson et al. 2024)
- testthat (Wickham et al. 2024d)
- randomForest (Liaw and Wiener 2024)
- ggplot2 (**ggplot2?**)
- maps (**maps?**)
- knitr (**knitr?**)
- here (**here?**)
- kableExtra (**kableExtra?**)

Measurement

The dataset, compiled by The Washington Post, translates real-world school shooting incidents into structured entries by aggregating information from news articles, open-source databases, law enforcement reports, and direct calls to schools. Only verified incidents, such as shootings during school hours or on school property, were included. Events like after-hours shootings, private suicides, or accidental discharges without other injuries were excluded (The Washington Post 2024b).

Summary Statistics & Relationship Between Variables

In this dataset, there are three variables that have relationship: $ Causalities = Killed + Injured $.

Table 1 summarizes the mean, median and standard deviation of casualties from school shooting events in US from 1999 to today. The standard deviation of 3.72 indicates that while most incidents have casualties close to the mean, there is a wide range of variability, with some incidents having significantly higher number of casualties.

Table 1: Summary Statistics for Casualties

Mean	Median	Standard Deviation
1.612981	1	3.720568

Summary Statistics for Casualties

Outcome variables

The outcome variable for this analysis is number of causalities for each state in 2025.

Figure 1 visualizes the number of school shooting incidents in US by state for all data since 1999. California, Texas, Florida and North Carolina all have more than 20 school shooting incidents in the past 25 years while other states has less than 20 school shootings.

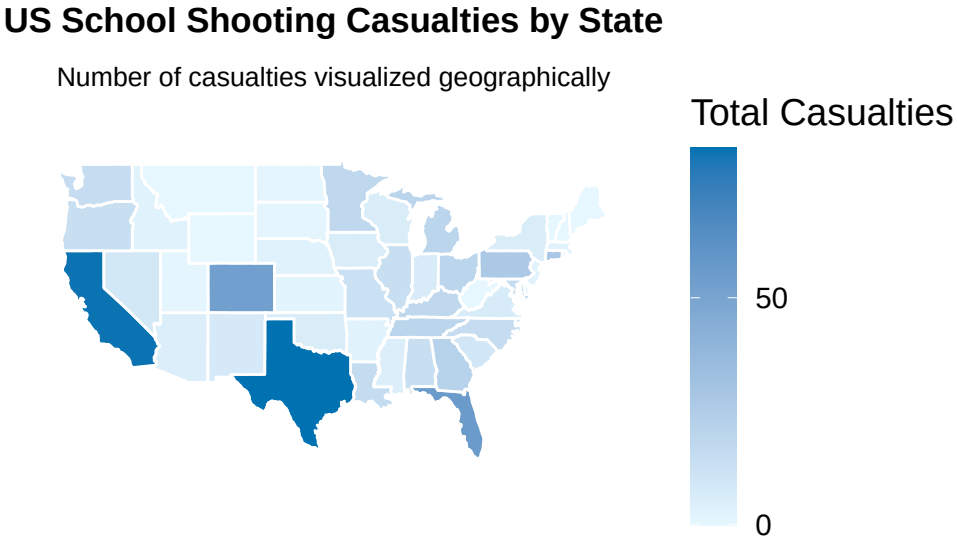

Figure 1: US School Shooting Casaulties by State

Predictor variables

Predictor variables in this analysis is state and year.

- State represents the state where the schools are located.
- Year represents the shooting year.

?@fig-Number_of_Incidents_Over_Time shows the number of school shooting incidents in US since 1999. The number of school shooting incidents almost doubled since 2020 compared with previous years.

Model

Model Set-Up

To model the total number of casualties resulting from school shooting incidents, a **Random Forest regression model** is employed. The model is designed to predict casualties based on state-level and temporal features derived from the dataset. The Random Forest algorithm is chosen for its robustness to non-linear relationships, its ability to handle interactions among predictors, and its ability to integrate categorical variables without creating dummy variables.

The mathematical representation of the model is: $C = f(X_1, X_2, ..., X_p) + \epsilon$

where:

- C: The total number of casualties (sum of fatalities and injuries) for a given year and state.
- $f(\cdot)$: The function estimated by the Random Forest model, representing the aggregate prediction from all decision trees.
- $X_1, X_2, ..., X_p$: Predictors, including:
  - X_1: Year (numeric), capturing temporal trends.
  - X_2: State (categorical), capturing regional effects.
- ϵ: The error term, representing unobserved factors affecting the number of casualties.

Model Justification

- **Year** is included to account for temporal trends in school shootings, reflecting potential increases or decreases over time.
- **State** is treated as a categorical variable to capture location-specific effects, such as differences in legislation, policing, or socioeconomic factors.

Number of trees is selected to be 300 to balance accuracy and prevent overfitting.

The choice of these features aligns with the data section, ensuring that variables with potential predictive value are incorporated without overfitting or introducing unnecessary complexity.

Assumptions and Limitations

- **Assumptions**:
  1. Casualties are conditionally independent given the predictors.
  2. The relationship between predictors and casualties can be approximated by the ensemble of decision trees.
  3. Data is representative of underlying trends and free of major sampling biases.
- **Limitations**:
  1. The model may not generalize well to future data if underlying trends shift dramatically (e.g., policy changes or societal events).
  2. Random Forests lack explicit interpretability compared to simpler models.

Software Implementation

The model is implemented in **R**, using the randomForest (**randomforest?**) package for training and evaluation.

Model Validation

- **Train/Test Splits**: The data is split into training (80%) and test (20%) sets. The model is trained on the former and validated on the latter to ensure generalizability.
- **Evaluation Metrics**:
  - **Root Mean Squared Error (RMSE)** and **Mean Absolute Error (MAE)** are calculated to assess prediction accuracy.
  - Final metrics:
    * RMSE: 1.491141 casualties
    * MAE: 1.096164 casualties

Alternative Models Considered

- **Linear Regression**:
  - Strengths: Simplicity and interpretability.
  - Weaknesses: Inadequate for capturing non-linear relationships and interactions present in the data.
- **Gradient Boosting Machines (GBM)**:
  - Strengths: Often more accurate due to its iterative learning.

Table 2: Summary Statistics for 2025 US School Shooting Casualties

Mean	Median	Standard Deviation
1.594306	1	1.910265

  - Weaknesses: Higher risk of overfitting, more computationally expensive and less intepretibility.

The Random Forest model was chosen over these alternatives for its balance of flexibility, robustness, and interpretability of feature importance.

Results

Since random forest is non-parametric, there is not parameter estimates. ????

Summary Statistics

Table 2 shows the summary statistics for predicted 2025 school shooting casualties. On average, each school shooting incident results in approximately 1.59 casualties. At least half of the incidents involve only one casualty. The standard deviation of 1.91 indicates that while many incidents cluster around the mean of 1.59, there is considerable spread, with some incidents resulting in significantly higher casualties.

Predicted Casualties by State

Using a Random Forest model, predictions for total casualties in each state for the year 2025 were generated. The predicted casualties were visualized geographically using a choropleth map in Figure 2. States such as California and Texas are predicted to experience the highest casualties, reflecting historical trends.

The predicted casualties for each state are summarized in Table 3. The table provides a clear and concise overview, sorted in descending order of predicted casualties to emphasize the most affected states.

Table 3: Predicted School Shooting Casualties by State for 2025

State	Predicted Casualties
georgia	4.6883274
california	4.6619440
pennsylvania	3.0609682
iowa	2.9120214
maryland	2.8487079
colorado	2.6481303
idaho	2.5716896
alabama	2.5343763
louisiana	2.4937956
nebraska	2.4329939
tennessee	2.3065005
washington	2.2872179
illinois	2.2528069
kansas	2.2381947
florida	2.1868345
south carolina	2.0988591
alaska	2.0804183
south dakota	2.0804183
virginia	2.0539586
missouri	1.9952304
nevada	1.8134811
arizona	1.7689940
hawaii	1.7424252
rhode island	1.7218750
kentucky	1.7020998
new jersey	1.6918049
arkansas	1.6550638
north dakota	1.6386594
district of columbia	1.5997442
new hampshire	1.5901607
delaware	1.5773379
north carolina	1.5707518
oklahoma	1.5016990
oregon	1.4668088
ohio	1.4616147
montana	1.4437321
mississippi	1.4289824
indiana	1.3932708
new york	1.3631713
new mexico	1.3572642
texas	1.3251580
wisconsin	1.2955882
connecticut	1.2038512
massachusetts	1.1329785
minnesota	0.9951597
utah	0.9079388

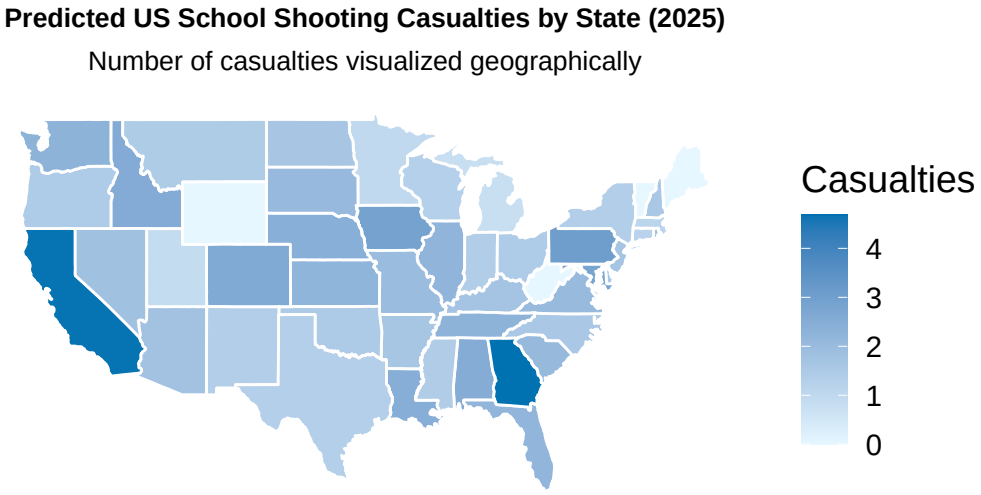

Figure 2: Predicted 2025 US School Shootings Numbers

Discussion

Actions to mitigate the impact of high-casualty incidents

The distribution of school shooting casualties reveals significant insights into the nature of these incidents. The mean number of casualties is higher than the median, indicating a positively skewed distribution. This suggests that while most school shooting incidents result in relatively few casualties, there are a small number of events that result in significantly higher numbers of casualties. These high-casualty incidents have a disproportionate impact and should be a focal point for policymakers. Understanding the causes and contexts of such incidents, such as the presence of high-density schools, specific socio-economic factors, or insufficient security measures, could help in designing targeted interventions. For instance, schools in high-risk areas could benefit from increased security personnel, improved mental health resources for students, and stricter firearm regulations. Government must prioritize strategies aimed at mitigating the impact of these high-casualty incidents to protect vulnerable populations.

Targeted Interventions for High-Risk States

The state-level predictions of casualties, as visualized in the map and table, provide actionable insights for policymakers at both the state and federal levels. States such as Georgia, California, and Pennsylvania are projected to experience higher numbers of casualties in 2025. This highlights the need for state-specific policies and resource allocation. For example, states with higher predicted casualties could implement tailored programs such as increasing mental health support in schools, conducting safety drills, and investing in school infrastructure to enhance safety measures. Additionally, identifying regions with consistently higher casualties can help

governments allocate resources more effectively and conduct targeted investigations to address root causes.

Policy Recommendations to Reduce Casualties

The map and table of predicted casualties highlight opportunities for governments to take preventative action. One key recommendation is to introduce universal background checks and stricter gun control measures in states with high predicted casualties. Furthermore, funding programs that promote community engagement and early intervention in at-risk populations could help reduce the likelihood of school shootings. Schools can also be equipped with better surveillance technology and emergency response systems to minimize casualties during incidents. At a broader level, creating national frameworks for school safety policies can ensure consistency and accountability in protecting students and staff across all states.

Weaknesses and next steps

While the Random Forest model provided reasonable predictions, it exhibited some limitations. The negative percentage of variance explained during initial iterations indicated that the model struggled to generalize with the available predictors. Although this was addressed by refining the data and removing outliers, the inclusion of only year and state as predictors limits the model's explanatory power. Additionally, the dataset's small size and imbalance across states might have impacted the model's ability to identify robust patterns.

The findings from this study underscore the importance of integrating predictive modeling with proactive policy-making. However, future work should focus on enhancing the dataset by incorporating additional predictors such as gun ownership rates, local crime statistics, and school-specific characteristics.

Additionally, examining the effectiveness of existing policies in high-risk areas could provide valuable insights into which interventions are most impactful.

Finally, collaborative efforts between federal agencies, state governments, and local communities will be essential to addressing the systemic issues underlying school shootings and reducing the number of casualties in future incidents.

Appendix

Additional data details

Model details

Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

Grolemund, Garrett, and Hadley Wickham. 2024. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Liaw, Andy, and Matthew Wiener. 2024. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. https://CRAN.R-project.org/package=randomForest.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal et al. 2024. *Arrow: Integration to 'Apache Arrow'*. https://CRAN.R-project.org/package=arrow.

The Washington Post. 2024a. "School Shootings Data." https://github.com/washingtonpost/data-school-shootings/blob/master/school-shootings-data.csv.

———. 2024b. "School Shootings Database." 2024. https://www.washingtonpost.com/education/interactive/school-shootings-database/.

Wickham, Hadley et al. 2024a. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

——— et al. 2024b. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

——— et al. 2024c. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

——— et al. 2024d. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

——— et al. 2024e. *The Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.