2025 US School Shooting Predictions and Solutions*

Leveraging Forecasting to Shape Policy and Protect Communities

Yun Chu

November 28, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

Paragraph of Estimand!

2 Data

2.1 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,

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

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
- maps
- knitr

2.2 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).

2.3 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

2.4 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.

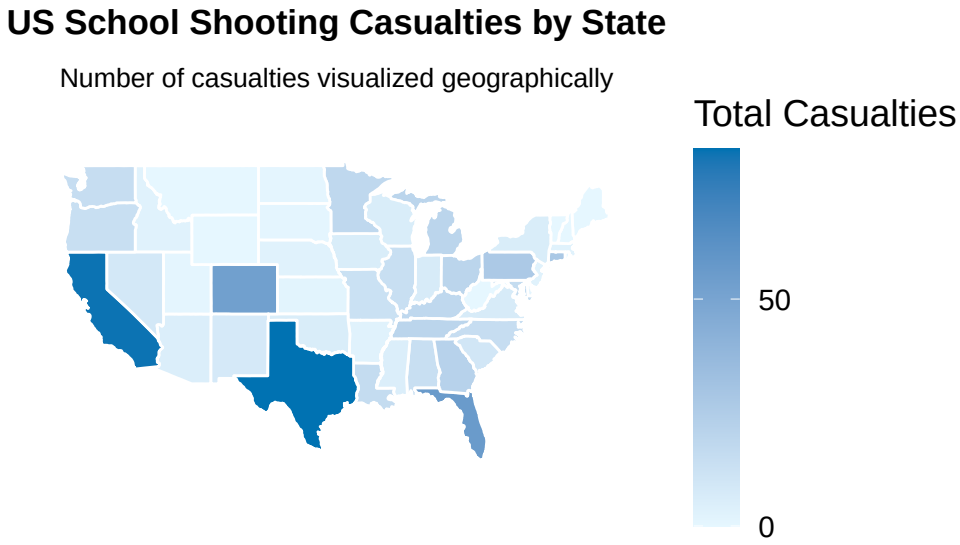

Figure 1: US School Shooting Casaulties by State

2.5 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.

3 Model

3.1 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.

3.2 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.

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.

3.2.1 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.

3.2.2 Software Implementation

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

3.2.3 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

3.2.4 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.
  - 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.

4 Results

Our results are summarized in Table **??**.

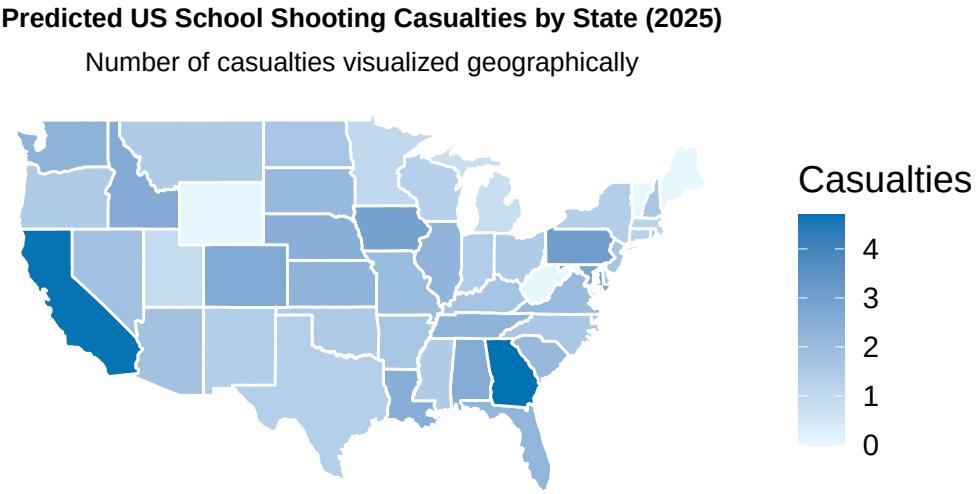

Figure 2

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 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

B.2 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.