2025 US School Shooting Predictions and Solutions*

Leveraging Forecasting to Shape Policy and Protect Communities

Yun Chu

December 2, 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.

1 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 (Jonson, 2017; RAND Corporation, 2015). 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 the existence of school shooting casualties, with the aim of informing policy and prevention efforts.

The estimand in this paper is to classify whether casualty exists in a school shooting incident in US. Using historical data from The Washington Post (The Washington Post 2024b), which provides a comprehensive record of school shootings, a Bayesian Logistic Regression model was developed to predict the existence of casualty based on shooting type and latitude of the school. The analysis includes summary statistics, spatial visualizations, and predictive modeling, providing actionable insights for policymakers. This study fills a critical gap in understanding the probability 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 2 discusses the data, including its sources, measurements, and key variables; Section 3 outlines the model used for prediction, including its justification, model validation is included in Section A; Section 4 presents the results, including summary statistics and spatial visualizations; and Section 5 concludes with a discussion of the findings, limitations, and recommendations for future research and policy.

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 and its students, 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 (The Washington Post 2024b). 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.

The statistical programming language R (R Core Team 2023) is used 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)
- modelsummary (Arel-Bundock 2023)
- ggplot2 (Wickham 2016)
- maps (Brownrigg et al. 2023)
- knitr (Xie 2023)
- here (Müller 2023)
- kableExtra (Zhu 2023)
- rstanarm (Gabriel A. Fonseca and Gelman 2023)

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: $causalities = killed + knjured$.

Table 1 summarizes the mean, median and standard deviation of casualties from school shooting events in US from 1999 till now. 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
1.61	1	3.72

Figure 1 illustrates the distribution of casualties across different shooting types. The "Targeted" and "Indiscriminate" categories exhibit significantly higher casualty counts compared to other types. As a result, these two types are combined into one class, while the remaining types form a separate class.

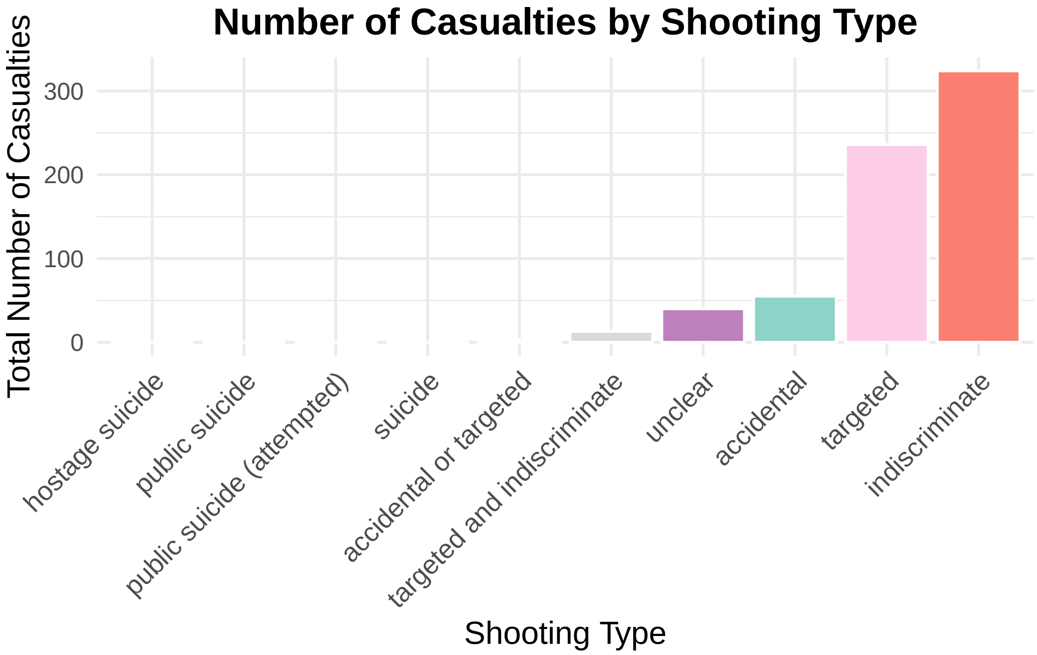


Figure 1: Distribution of Casualty By Shooting Type

2.4 Outcome variables

The outcome variable for this analysis is causalities for each state in 2025. In the model, we classify the existence of casualties.

Figure 2 illustrates the distribution of casualties. The data is positively skewed, with the majority of cases involving fewer than 7 casualties, though there are a few school shooting incidents with exceptionally high casualty counts.

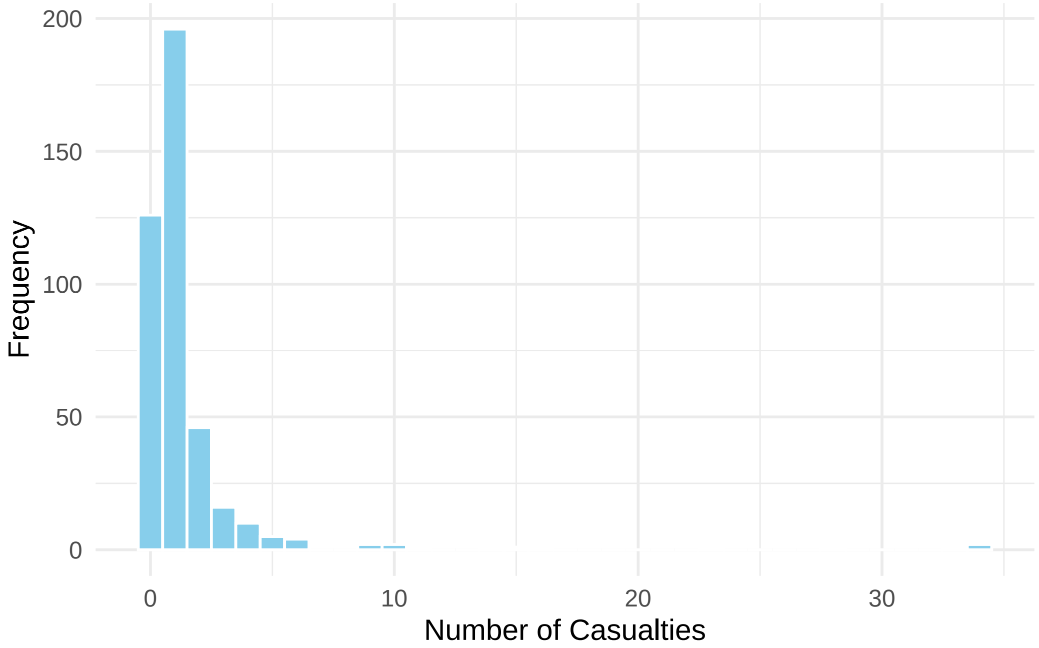


Figure 2: Distribution of Casualties

Figure 3 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.

2.5 Predictor variables

Predictor variables in this analysis is Shooting Category and Latitude.

- Shooting Category is a binary variable derived from Shooting Type with “Targeted” and “Indiscriminate” as one type while the rest is classified as the second type. Shooting Type represents the type of shootings with 5 categories: accidental, hostage suicide, indiscriminate, public suicide, targeted, and unclear. There are also some rows with a combination of these 5 types, for example, “accidental or targeted” and “targeted and indiscriminate”.
- Latitude represents the latitude of the school.

Figure 4 presents a pie chart depicting the proportions of various shooting types. Targeted shootings account for 56% of all school shooting incidents.

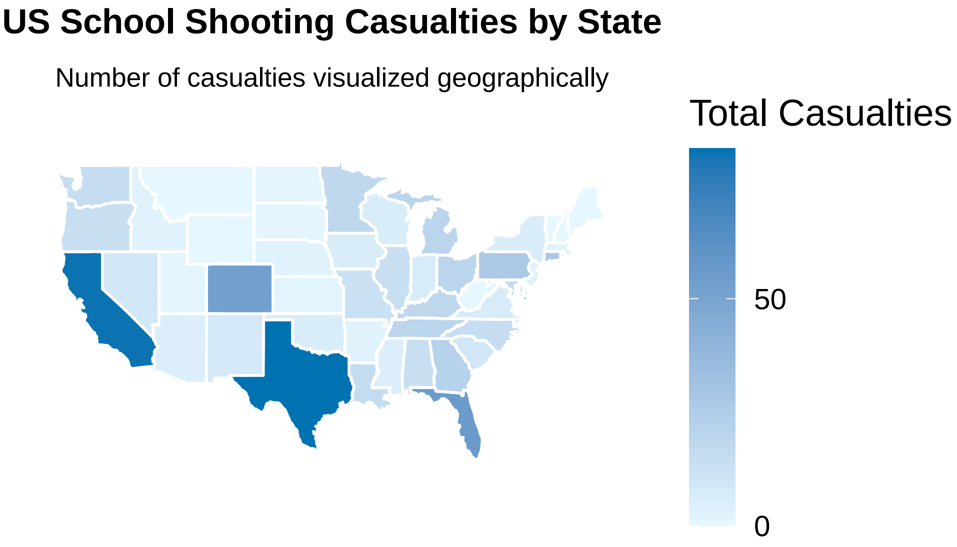

US School Shooting Casualties by State
Number of casualties visualized geographically
Total Casualties
50
0

Figure 3: US School Shooting Casaulties by State

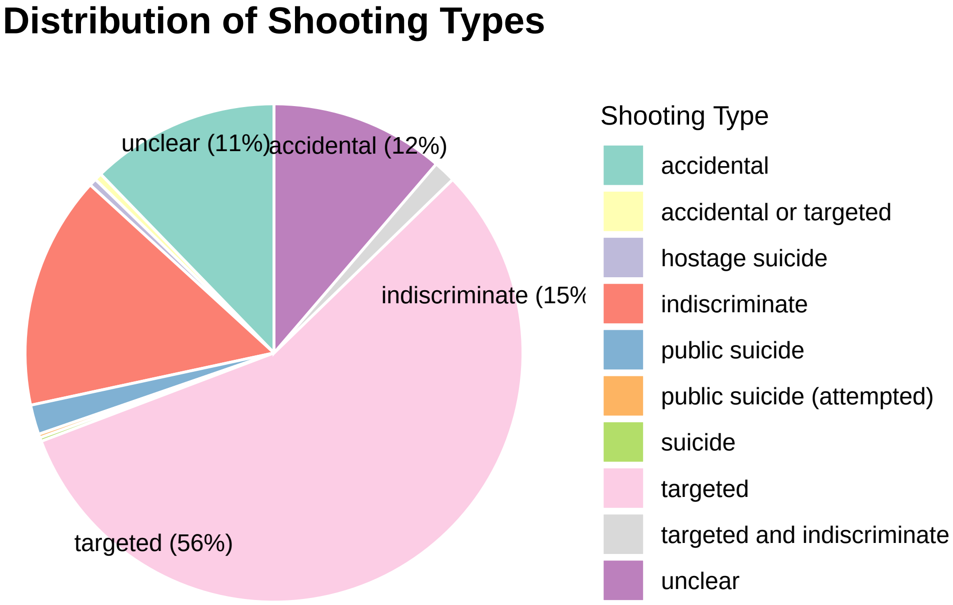


Figure 4: Shooting Type Distribution

Figure 5 shows the distribution of latitude values of schools in school shooting events, ranging from approximately 20 to 50 degrees. Most events occur between 30 and 40 degrees, with a peak near 35 degrees. There are few occurrences at higher latitudes (above 45 degrees).

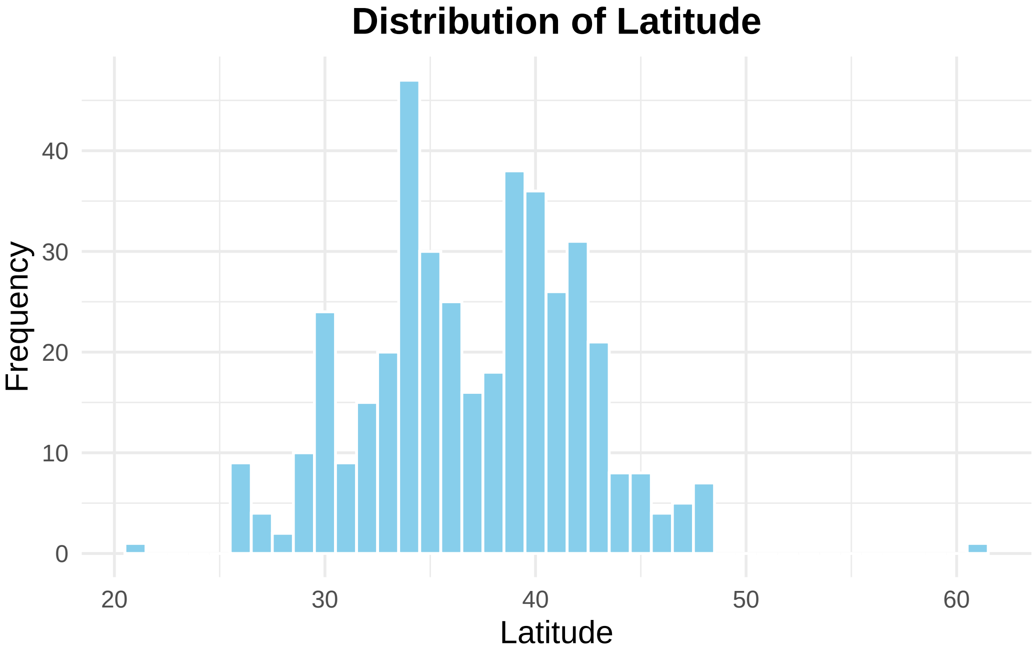


Figure 5: Distribution of Latitude of Schools in School Shooting Events

3 Model

3.1 Model Set-Up

The analysis employs a Bayesian Logistic Regression Model to predict whether a school shooting incident results in casualties. The binary response variable (`ifcasualty`) is defined as 1 if there are any casualties (injuries or fatalities) and 0 otherwise. The predictors include the **latitude** of the school where the incident occurs and a simplified categorical variable, **shooting category**, which distinguishes between "indiscriminate" or "targeted" shootings and "other" types.

3.1.1 Mathematical Formulation

The model estimates the probability of a casualty using the following formulation:

$$P(C = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$
$$\beta_0 \sim \text{Normal}(0, 5)$$
$$\beta_1, \beta_2 \sim \text{Normal}(0, 2.5)$$

where:

- $P(C = 1 \mid X)$: Probability of at least one casualty occurring in an incident.
- β_0: Intercept term.
- β_1: Coefficient for the shooting category, capturing the difference between "indiscriminate or targeted" and "other" types.
- β_2: Coefficient for latitude, accounting for geographic trends.
- X_1: Shooting category (0 = "indiscriminate or targeted", 1 = "other").
- X_2: Latitude of the school where the incident occurs (continuous).

The priors for the Bayesian logistic regression model are specified as follows:

- **Intercept (β_0)**: Reflects uncertainty about the baseline log-odds of casualties.
- **Coefficients (β_1, β_2)**: Weakly informative priors that regularize the model and prevent overfitting.

These priors were chosen to represent plausible parameter values while maintaining model stability.

The Bayesian logistic regression model is implemented in **R** using the `rstanarm` package. The final model is saved as an RDS file for reproducibility and further analysis.

3.1.2 Model Justification

The Bayesian logistic regression model is well-suited for this problem as it is specifically designed to predict binary outcomes, such as whether casualties occur, without requiring transformation of skewed response variables. Additionally, the Bayesian framework provides probabilistic predictions, allows for the incorporation of prior knowledge, and offers a robust quantification of uncertainty in parameter estimates.

The shooting types "Targeted" and "Indiscriminate" were combined into a binary variable alongside other shooting types, as exploratory data analysis in Figure 1 revealed that these two categories are strongly associated with a higher number of casualties.

Latitude is included as a predictor to account for geographic variability that might affect casualty likelihood. Longitude was evaluated during model selection but excluded due to its lack of statistical significance.

3.2 Model Limitations

The predictive performance of the Bayesian logistic regression model may be limited under certain circumstances. One key limitation is that the model's predictions might not generalize well if there are significant changes in societal or policy conditions, as these factors could alter the relationships between predictors and the outcome. Additionally, the model's reliability is sensitive to the quality of both the prior distributions and the dataset. Poorly chosen priors or biased and incomplete data can lead to inaccurate predictions and reduced robustness of the model.

3.3 Model Validation

The Bayesian logistic regression model is validated through several diagnostics. In Figure 7a, the posterior predictive check shows alignment between the posterior predictive distribution (y_{rep}) and the observed data (y), indicating a good fit. Figure 7b shows that posterior estimates are narrower and shifted compared to priors, reflecting the influence of the data. Additionally, Figure 8a confirms well-mixed and stationary MCMC chains, while Figure 8b shows $\hat{R} \approx 1.00$ for all parameters, verifying convergence and reliability.

3.4 Alternative Models Considered

Several models were evaluated during the development process:

1. **Linear Logistic Regression**:
   - **Strengths**: Simplicity and interpretability.
   - **Weaknesses**: Cannot incorporate prior knowledge or provide uncertainty quantification.
2. **Random Forest**:
   - **Strengths**: Handles complex, non-linear interactions.
   - **Weaknesses**: Limited interpretability and probabilistic outputs.

The Bayesian logistic regression model was chosen for its ability to provide interpretable, probabilistic predictions while quantifying uncertainty and incorporating prior information.

4 Results

The model results are summarized in Table 2.

Table 2: Explanatory model of the Existence of Casualty in Each School Shootig Event based on Shooting Type and Latitude

	Casualty Occurrence
(Intercept)	2.68
	(1.08, 4.31)
shooting_categoryother	−0.63
	(−1.07, −0.18)
lat	−0.04
	(−0.09, 0.00)
Num.Obs.	415
ELPD	−252.2
ELPD s.e.	8.4
LOOIC	504.5
LOOIC s.e.	16.8
WAIC	504.5

4.1 Intercept

The intercept estimate is **2.6798** (95% credible interval: **1.0774 to 4.3091**), representing the baseline log-odds of a casualty occurring when predictors are at their reference levels. This corresponds to a high baseline probability, with the credible interval excluding zero, indicating strong evidence for a nonzero probability of casualties under these conditions.

4.2 Shooting Type

The coefficient for the "other" shooting category is **-0.6336** (95% credible interval: **-1.0735 to -0.1784**), suggesting a lower likelihood of casualties compared to the "indiscriminate" or "targeted" categories.

4.3 Latitude

The latitude coefficient is **-0.0443** (95% credible interval: **-0.0876 to -0.0021**), indicating a slight but significant decrease in the probability of casualties with increasing latitude. This suggests potential geographic trends influencing casualty outcomes, such as differences in population density, response capabilities, or regional policies.

4.4 Predicted Existence of Casualties by State

Using a Bayesian Logistic regression model, predictions for the existence of casualties in each state were generated. The predicted casualties were visualized geographically using a choropleth map in Figure 6.

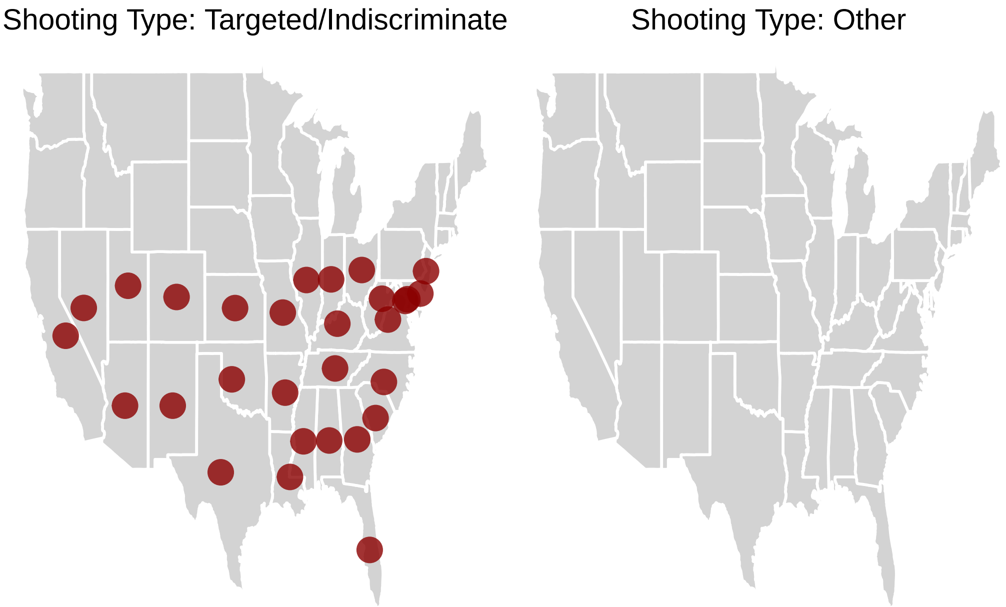


Figure 6: Existence of Casualties in US School Shootings Prediction

5 Discussion

5.1 Contributions of This Study

This paper utilizes a Bayesian Logistic Regression model to predict the likelihood of casualties in school shootings, focusing on the impact of shooting type and geographic latitude. The findings demonstrate that “Targeted” and “Indiscriminate” shootings significantly increase casualty probabilities, necessitating their classification into a unified predictive category. This categorization improves the model’s predictive accuracy and highlights the distinct risk these shooting types pose. Furthermore, the model generates spatial predictions at the state level, identifying regions with elevated casualty risks. These predictions provide policymakers with concrete data to prioritize interventions and allocate resources effectively, especially in high-risk areas.

5.2 Policy Implications

5.2.1 Geographic Targeting of Resources

The geographic analysis in this study reveals significant disparities in casualty risks, with southern states showing higher probabilities of school shootings resulting in casualties. This finding emphasizes the need for tailored interventions in these regions. For example, policymakers can allocate funding to implement enhanced physical security measures in schools, such as metal detectors, reinforced entry points, and surveillance systems. Additionally, improving emergency response systems, including faster deployment of law enforcement and medical personnel, can mitigate the impact of shootings in high-risk areas.

5.2.2 Focus on High-Risk Shooting Types

The model's findings confirm that "Targeted" and "Indiscriminate" shootings are strongly associated with higher casualty rates. Policymakers can design prevention strategies specifically aimed at these shooting types. For "Targeted" shootings, schools can adopt proactive threat assessment programs to identify and provide support to at-risk individuals. For "Indiscriminate" shootings, community-level initiatives addressing factors such as mental health, social isolation, and access to firearms can reduce the likelihood of such incidents.

5.2.3 Strengthening Gun Control Measures

Given the role of firearms in school shootings, this study highlights the importance of stricter gun control measures. Policies such as universal background checks, red flag laws, and safe firearm storage requirements can reduce unauthorized access to firearms and lower casualty risks. These measures can be particularly impactful in regions identified as having elevated risks, complementing other targeted interventions.

5.2.4 Investment in Mental Health Resources

The findings also point to the importance of addressing underlying causes, such as untreated mental health conditions, that may contribute to school shootings. Policymakers can expand access to mental health services by increasing funding for school counselors and psychologists, creating community-based support programs, and implementing mental health education initiatives. These efforts can reduce risk factors while fostering a supportive environment for students.

5.2.5 Ongoing Policy Evaluation

Effective policymaking requires regular assessment of interventions. This study provides a framework for evaluating the impact of existing measures, such as increased school security or stricter firearm regulations. Policymakers can use predictive models to monitor changes in casualty risks and adapt strategies based on emerging data. For example, periodic updates to the model with new data can refine predictions and improve policy responsiveness.

5.3 Limitations of the Analysis

The reliability of the model depends on the availability and quality of the underlying data and priors. Changes in societal conditions or policy environments may reduce the model's predictive power over time. Additionally, the analysis does not account for several important contextual factors, such as individual school characteristics, socio-economic disparities, and differences in firearm accessibility across states. Including these factors could improve the model's ability to capture the complexity of school shooting dynamics.

5.4 Recommendations for Future Research

5.4.1 Expanding the Predictive Framework

Future studies should incorporate additional predictors to enhance the model's scope and accuracy. Factors such as school size, local economic conditions, and access to firearms could provide a more detailed understanding of the drivers of casualties. Incorporating longitudinal data could also help capture temporal trends and the impact of policy changes over time.

5.4.2 Evaluating Policy Effectiveness

To ensure that interventions are impactful, future research should evaluate the effectiveness of existing policies. For example, studies could assess whether stricter gun control laws or increased mental health funding significantly reduce the likelihood or severity of school shootings. Policymakers could use these insights to refine strategies and develop evidence-based policies.

5.4.3 Encouraging Interdisciplinary Collaboration

Addressing the complex issue of school shootings requires collaboration across disciplines. Future research should bring together experts from education, mental health, public policy, and law enforcement to design holistic interventions. Collaborative efforts can also facilitate data sharing, improve predictive modeling, and create a more comprehensive understanding of how to reduce casualties effectively.

A Appendix

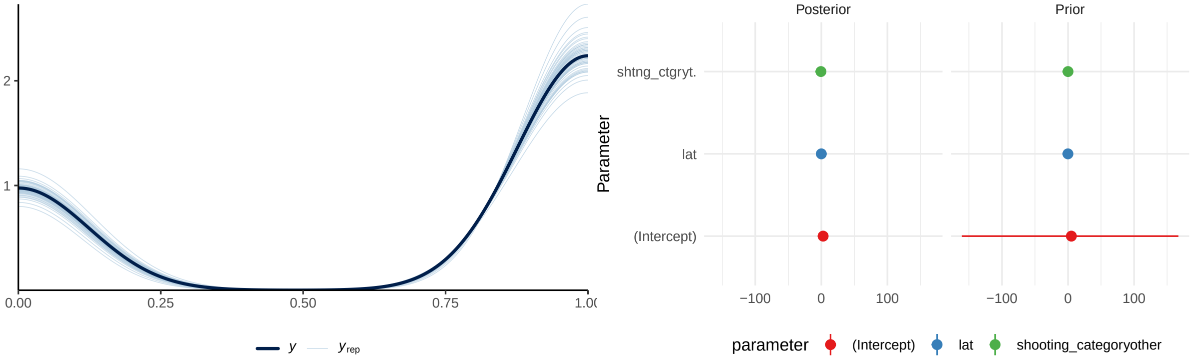
A.1 Additional data details

A.2 Model details

A.3 Posterior predictive check

In Figure 7a, we implement a posterior predictive check. This shows that the posterior predictive distribution (y_{rep}) aligns well with the observed data (y), indicating that the model provides a reasonable fit to the data and captures its underlying structure effectively.

In Figure 7b, we compare the posterior with the prior distributions. This shows that the posterior estimates for all parameters ($Intercept$, lat, and $shooting\_categoryother$ are informed by the data, as the posterior distributions are narrower and shifted relative to the priors, reflecting updated beliefs based on the observed evidence.



(a) Posterior prediction check

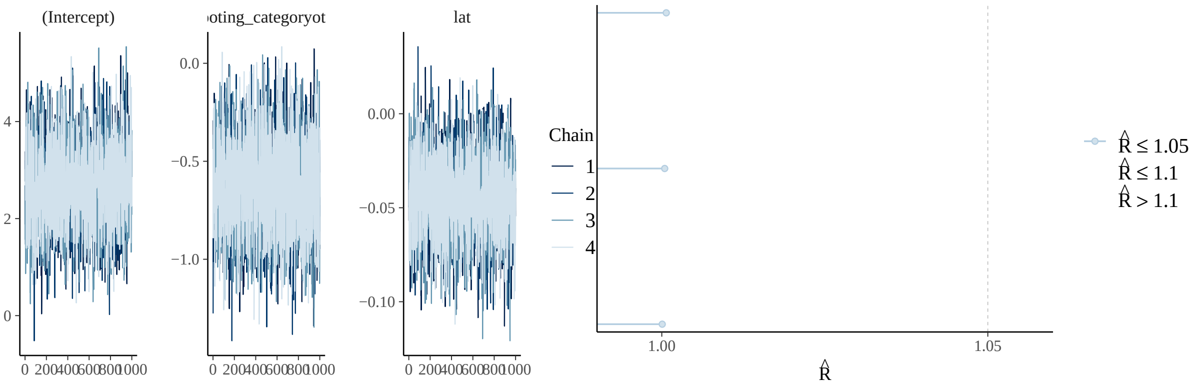
(b) Comparing the posterior with the prior

Figure 7: Examining how the model fits, and is affected by, the data

A.4 Diagnostics

Figure 8a is a trace plot. It shows the sampling behavior of the MCMC algorithm for the model parameters (Intercept, shooting_categoryother, and lat) across four chains. The chains are well-mixed, stationary, and overlap significantly, indicating good convergence and effective exploration of the posterior distributions. It suggests that the model's parameters were sampled effectively, with no immediate signs of convergence issues.

Figure 8b is a Rhat plot. This $\hat{R}$ plot shows that all parameters have $\hat{R} \approx 1.00$, indicating excellent convergence of the MCMC chains. The results suggest that the chains are well-mixed, and the posterior distributions can be considered reliable for interpretation.



(a) Trace plot

(b) Rhat plot

Figure 8: Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data*. https://vincentarelbundock.github.io/modelsummary/.

Brownrigg, Ray, Thomas P Minka, Robert A Becker, Allan R Wilks, Original S code by Richard A Becker, and Allan R Wilks. 2023. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Gabriel A. Fonseca, Ben Goodrich, Jonah Gabry, and Andrew Gelman. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

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

Müller, Kirill. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org/.

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.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. Yihui Xie. https://yihui.org/knitr/.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.