IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

Predictive Modeling of Gun Violence Using Machine Learning: Understanding the Role of Demographic and Socioeconomic Factors at the County Level

Rohan Singhal

6901 Coit Rd, Frisco, TX 75035, USA; singhal.rohan06@gmail.com
Mentor: Mr. Brian Lewis

ABSTRACT: In a society plagued by the implications of gun violence, the need for effective risk assessment measures has become evident. This project addresses the challenge of predicting gun violence risks at the county level in the United States by harnessing the power of machine learning-based predictive modeling. We hypothesize that demographic and socioeconomic factors, such as unemployment and poverty rates in 2021, influence the likelihood of gun violence in various regions. Through an intricate analysis of multifaceted datasets encompassing crime rates, socioeconomic indicators, and geographic characteristics, this research seeks to uncover the extent to which these factors impact the risk of gun violence at the county level. Analyzing 2021 data, the research identifies counties with specific poverty-related characteristics that significantly elevate the risk of shootings. The study's use of decision trees provides a glimpse into the predictive model's classification process. By decoding the complex dynamics of gun violence, this research opens new avenues for understanding and addressing the factors contributing to shootings, ultimately guiding the way toward a safer society.

KEYWORDS: Robotics and Intelligent Machines, Machine Learning, Gun Violence, Poverty, Demographic Factors.

Introduction

Gun violence is a pressing issue in modern society, particularly in the United States, where devastating mass shootings have prompted urgent discussions on prevention.[1] Despite the focus on gun control policies, understanding the multifaceted factors underlying gun violence, especially at the county level, remains a critical gap.[2] This research uses advanced predictive modeling techniques to analyze gun violence risk at the county level, highlighting the intricate relationship between demographic, socioeconomic, and geographic factors that influence these occurrences.

Gun violence poses a complex challenge with no single cause. It is influenced by various factors, including social, economic, and geographic dynamics. By recognizing this complexity, this research delves into county-level data to unveil patterns shaping gun violence in the United States. Leveraging comprehensive datasets, including crime rates, poverty, unemployment, and educational indicators, this analysis seeks to unravel the nuanced relationship between socioeconomic status and gun violence risks in specific regions.

Our hypothesis posits that demographic and socioeconomic factors, among other variables, contribute significantly to varying levels of gun violence risks across counties. Comprehensive datasets are vital to exploring this intricate relationship, enabling us to move beyond reactive measures and adopt proactive, data-driven approaches. This research contributes to the growing knowledge surrounding gun violence by employing cutting-edge technology and in-depth analysis. It informs evidence-based policies and interventions, paving the way for safer communities and bolstering social security.

Literature Review:

The prediction of gun violence, particularly school shootings, is a topic of great concern and research interest. Understanding the underlying factors and patterns contributing to these incidents is crucial for prevention and intervention efforts. This literature review explores recent research that delves into the intricate task of predicting gun violence, focusing on community-level factors, machine-learning approaches, and dynamic factor models.3

Blair T. Johnson's study "Community-level factors and incidence of gun violence in the United States, 2014-2017" investigates the role of community-level factors in the incidence of gun violence. The research examines data from 2014 to 2017 and identifies variables at the community level that may contribute to gun violence. Johnson's work found that the extent to which counties are urban was the most robust predictor of both gun violence incidents and casualty rates. Similarly, places characterized by greater income disparity were also more likely to experience higher gun violence rates, especially when high income was paired with high poverty.[4] These findings highlight the complex relationship between urbanization, income disparity, and gun violence, providing valuable insights for future predictive models.

Dana E. Goin's study "Predictors of firearm violence in urban communities: A machine-learning approach" employs machine-learning methodologies to predict firearm violence within urban communities. By analyzing a range of variables, including demographic factors and neighborhood characteristics, Goin's research found that certain variables, such as neighborhood poverty rates and educational attainment, sig-

DOI: 10.36838/v6i8.3

nificantly influence firearm violence.[5] This machine learning approach provides a data-driven foundation for developing effective strategies to address gun violence in urban settings.

Salvador Ramallo's research "A dynamic factor model to predict homicides with a firearm in the United States" introduces a dynamic factor model to predict homicides involving firearms. This innovative approach combines dynamic modeling techniques with comprehensive data to forecast firearm-related homicides. Ramallo's work reveals the temporal dynamics of gun violence, indicating that specific factors, such as firearm sales and legislative changes, impact the occurrence of firearm-related homicides.[6] This dynamic factor model offers a novel perspective for predictive modeling in this critical area, emphasizing the importance of considering time-varying factors in gun violence prediction.

These studies represent a diverse range of research endeavors to advance our understanding of gun violence prediction. They underscore the importance of community-level factors, machine learning, and dynamic modeling in enhancing predictive accuracy. The findings from these studies highlight various factors, including urbanization, income disparity, neighborhood poverty rates, educational attainment, firearm sales, and legislative changes, as significant predictors of gun violence. Building upon the foundations laid by these scholars, this project aims to contribute to the ongoing discourse on predictive modeling for gun violence by analyzing how various measures of poverty and income disparity play a specifically predictive role in gun violence likelihood. This paper demonstrates more granularity and nuance into which specific measures of poverty and income have strong predictive power in gun violence outcomes across the United States.

Methods

Data Collection and Selection:

This project's foundation lies in meticulously collecting and curating diverse datasets that capture the various dimensions of gun violence and potential determinants. Multiple sources were explored to acquire relevant data, ensuring a comprehensive representation of the factors influencing gun violence at the county level. The key data sources encompassed the Center for Homeland Defense and Security (CHDS) shooting incidents dataset, the Mother Jones mass shootings dataset, the Integrated Public Use Microdata Series (IPUMS) census data, Data.gov datasets,[7] Federal Communications Commission datasets,[27] and county-level crime statistics from the Federal Bureau of Investigation (FBI) Uniform Crime Reporting (UCR) program. The selection of these datasets was driven by their comprehensive coverage of variables related to gun violence, socioeconomic indicators, and demographic factors.

Data Preprocessing:

The selected datasets were carefully cleaned and transformed to ensure their compatibility for analysis. Python emerged as the programming language of choice due to its versatility and extensive libraries for data manipulation.[8] The CHDS and Mother Jones datasets were merged, leveraging the city, state, and date variables to create a unified incident timeline. The IPUMS census data was integrated using the Federal Information Processing Standards (FIPS) code, a unique identifier for each county. This step facilitated the alignment of socioeconomic and demographic information with gun violence incidents.

Feature Selection:

The complexity of the relationship between gun violence risks and the myriad of predictor variables necessitated a thoughtful approach to feature selection. Key demographic variables, such as population density,[9] education,[10] and racial composition, were chosen due to their potential influence on crime rates. Socioeconomic factors, including poverty and unemployment rates for 2021,[12] were also included, reflecting their strong connection to community well-being and crime. These features were augmented with crime data,[12,13] encompassing various offense categories, allowing for a comprehensive assessment of the crime landscape.

Geographic Integration:

An essential component of the analysis involved integrating geographic information with the datasets. The Google Geocoding API and Google Maps Geolocation Services were utilized to geocode the location information provided in the datasets and derive latitude and longitude coordinates. These coordinates were then employed to query the API, yielding the FIPS code, which encapsulates the geographical identifier at the county level.[14] This enabled the linkage of gun violence incidents to specific counties, facilitating spatial analysis.

Long-Wide Data Conversion:

To extract temporal insights, the data underwent a conversion from a wide to a long format.[15,16] This transformation allowed us to extract the year from variables related to unemployment and poverty rates, creating a separate year column. Subsequently, the data was transformed back to its original wide format, now enriched with this temporal information. This approach provided a granular temporal dimension, enabling a closer examination of trends and patterns over time.

Observational Level:

The analysis is conducted at the county level, with each county-year combination represented by a unique observation unit. This approach facilitates a granular examination of how different variables contribute to gun violence risks across diverse counties over time.[17-20] The merged dataset encompasses a comprehensive timeline of gun violence incidents,[12,13] demographic indicators,[20] socioeconomic metrics,[21,22] and crime rates, creating an integrated resource for analysis.

By intricately manipulating and integrating diverse datasets, harnessing geographic insights through the Google Geocoding API and Google Maps Geolocation Services, and employing advanced Python programming techniques, this methodology provides a robust foundation for the subsequent stages of analysis.

Ten-fold Cross Validation and Class Imbalance:

As the analysis progressed, an essential aspect of model selection and evaluation was addressing the challenge of class imbalance with the gun violence dataset. The logistic regression model, initially employed, revealed an unexpected pattern in its predictions. While the model appeared accurate when predicting the absence of a shooting (coded as 0), its performance significantly deteriorated when predicting the occurrence of a shooting (coded as 1). This disparity stemmed from the class imbalance present in the dataset, where instances of shootings were notably less frequent compared to non-shooting instances.

Class imbalance refers to an unequal distribution of classes in a dataset, which can lead to skewed model predictions and compromised accuracy. A transition was made to a tree-based model, specifically a Random Forest classifier, to overcome this issue. This decision was motivated by their inherent capacity to handle nonlinear relationships and adapt to imbalanced datasets.

To refine the model's performance, ten-fold cross-validation emerged as a feasible technique. Ten-fold cross-validation involved splitting the dataset into ten subsets or "folds". The model is trained on nine of these folds and tested on the remaining fold. This process is repeated ten times, ensuring that each fold serves as the testing set once.[23] The average performance across the ten iterations provides a more robust estimate of the model's generalization ability.

By implementing ten-fold cross-validation, the performance of the Random Forest model was evaluated across different subsets of the data. This technique effectively addressed the issue of class imbalance and offered insights into the model's consistency in predicting gun violence incidents. The outcome of this validation process proved promising, demonstrating a marked improvement in the model's accuracy and predictive power compared to the initial logistic regression attempt.

To evaluate the models' performance, various metrics were employed. The logistic regression model used metrics such as accuracy, precision, and F1 score, specifically weighted F1 and macro F1.[24] These metrics provided a nuanced understanding of the model's predictive power. Similar metrics were used for the Random Forest model, with a particular focus on accuracy.

Result and Discussion

In our analysis of predictors for shootings in 2021, poverty-related variables were the most significant. The feature importance plot shown below highlighted three key factors: counties with wider variations in total poverty (higher upper bounds), consistently high child poverty rates (high lower bounds), and more related children aged 5 to 17 in family poverty faced a significantly elevated risk of shootings.[25] Conversely, the overall percentage of total poverty and its confidence intervals displayed weaker associations with shooting likelihood.

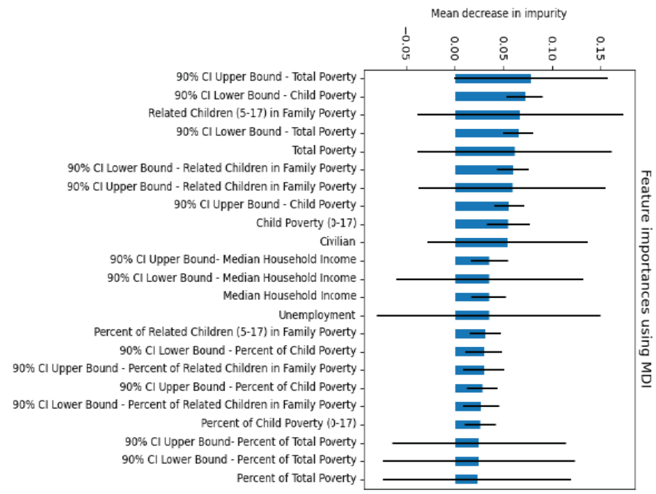

Figure 1: Feature importance plot showing different variables and their relative contribution to gun violence risk.

Based on Figure 1, it is evident that the total number of people in poverty is the most influential factor, significantly impacting the likelihood of shootings occurring at the county level. Surprisingly, the percentage of people in poverty appeared to be the least influential, suggesting a nuanced relationship between poverty rates and gun violence. This observation indicates that the sheer size of a county's population in poverty[20,25] plays a crucial role in influencing the occurrence of shootings. This finding draws attention to the significance of understanding not only the prevalence but also the demographic scale of poverty within a community, highlighting the multifaceted nature of gun violence risk.

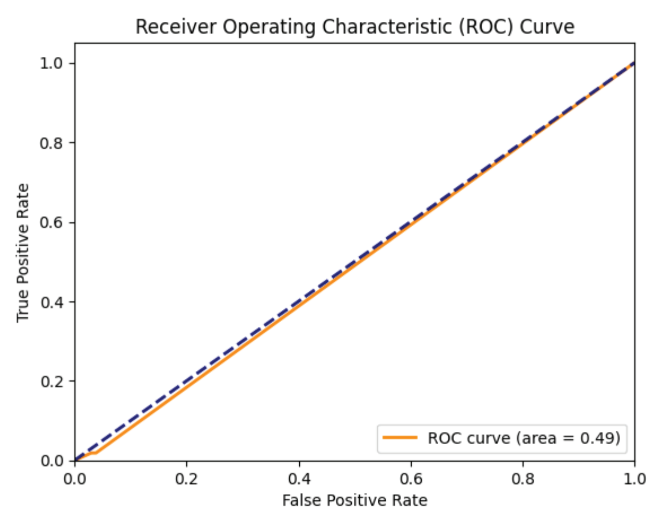

Figure 2: ROC Curve showing the performance of the logistic regression model.

During the exploratory phase of this research, we initially experimented with a logistic regression model to predict gun violence occurrences at the county level. However, the results revealed that logistic regression was not suitable for our dataset due to its limited predictive power. Based on Figure 2, the ROC curve for the logistic regression model showed an area

DOI: 10.36838/v6i8.3

under the curve (AUC) score of 0.49, which is close to random chance (0.50) and indicates that the model's ability to distinguish between shootings and non-shootings was no better than flipping a coin. This lack of predictive power led us to explore alternative machine learning techniques that could handle the complexity and nonlinear relationships within the data. As a result, we transitioned to using a random forest classifier, which demonstrated far better performance in predicting gun violence risk. The decision to use random forest over logistic regression was driven by the random forest model's ability to capture complex interactions and nonlinearities in the data, resulting in a more accurate and robust predictive model.

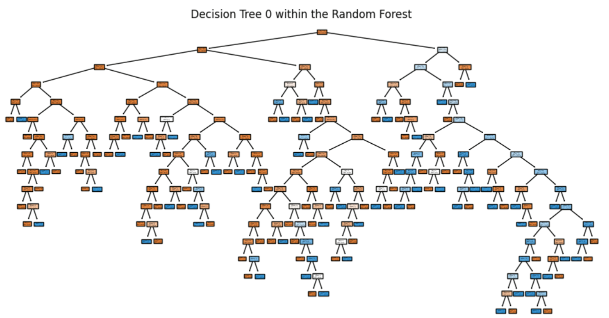

Figure 3a: Structural overview of the decision tree and its classification process. Nodes represent decision points determined by specific features.

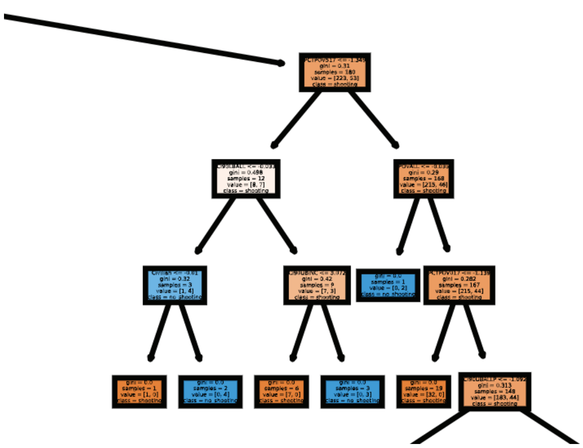

Figure 3b: The zoomed-in image of the decision tree provides a detailed view of the specific logic and classification criteria within a selected segment of the entire decision tree (Figure 3a).

We used decision trees to help visualize the model's decision logic to enhance our understanding. Based on Figure 3a, although the tree shown above is one example out of many, it illustrates how the model classifies counties based on the aforementioned factors. Each box in the tree represents a decision node where the data is partitioned based on specific conditions, ultimately categorizing the data as either shooting or no shooting, coded as 0 or 1, respectively. While this decision tree is simplified, it accurately captures how the model classifies counties.

The first variable in each box corresponds to one of the features from our dataset. For instance, it could be a feature like "Total Poverty" or "Unemployment Rate" measured at the county level. The condition associated with this variable is usually a threshold or range (e.g., less than or equal to a specific value) that determines how the data is split at that node.

The Gini coefficient is a measure of impurity or uncertainty within a node. It quantifies the probability of misclassifying a randomly chosen sample, with lower values indicating higher purity or homogeneity in terms of the target variable (shooting vs. non-shooting).

The "samples" variable indicates the number of samples (data points or instances) that fall into the specified condition at that node. It provides insight into the distribution of data and the volume of instances affected by the split condition.

The "value" variable within brackets represents the distribution of classes (shooting and non-shooting) within the samples at that node. For example, if the value is [300, 100], it means there are 300 instances classified as non-shooting and 100 instances classified as shooting.

Based on the split condition, Gini coefficient, and distribution of classes (value), the decision tree algorithm determines the class (shooting or non-shooting) for each node. The majority class or the class with higher representation within the node is often assigned as the predicted class for that subset of data. The decision tree's hierarchical structure allows it to recursively split the data based on the most informative features and conditions, gradually creating branches that lead to the final classification decisions. By following the path from the root node to the leaf nodes, we can trace how the model makes decisions at each step, ultimately predicting whether a county is likely to experience gun violence or not based on the given features and conditions.

Our study emphasizes the pivotal role of poverty-related factors, especially those with wider variations and higher lower bounds, in predicting shootings. These insights offer a better understanding of the dynamics of shooting incidents, which also helps mitigate their occurrence.

The random forest model's performance was evaluated using ten-fold cross-validation. Across the folds, the accuracy was consistently high, ranging from 0.91 to 0.95, and the average accuracy across the 10 folds was 0.93. These accuracy scores highlight the model's ability to correctly predict county-level gun violence risk across diverse datasets.

The model's effectiveness was further measured using the weighted F1 score, which accounts for both precision and recall. The F1 weighted scores ranged from 0.90 to 0.95 across the folds, and the average weighted F1 score across 10 folds was 0.92. This indicates the model's balanced performance in predicting both positive and negative instances.

The model's performance was also assessed using the macro F1 score, which considers the balance between precision and recall for each class. The F1 macro scores ranged from 0.55 to 0.81 across the folds, with an average score of 0.67. These results demonstrate the model's ability to generalize well to diverse data and maintain consistent performance across different subsets.

Moreover, the analysis of feature importance revealed key factors influencing gun violence risk. Poverty-related variables, such as the 90% Confidence Interval Upper Bound for To-

DOI: 10.36838/v6i8.3

tal Poverty, exhibited the most substantial impact. Counties with higher upper bounds of total poverty were more likely to experience gun violence incidents. This insight aligns with the broader understanding that socioeconomic disparities[26] significantly contribute to the occurrence of gun violence at the county level.

Conclusion

From this analysis, poverty emerges as a critical determinant of shootings in 2021. Counties characterized by higher upper bounds of total poverty, elevated lower bounds of child poverty, and a greater number of related children aged 5 to 17 in family poverty face an increased risk of shootings. This highlights the relationship between socioeconomic factors and gun violence, emphasizing the need for approaches that address these poverty-related disparities. While this study provides valuable insights, it also highlights the limitations of focusing solely on one year of data. Future research should encompass broader temporal and spatial scopes, incorporating additional variables such as educational attainment, mental health statistics, and gun ownership sales to gain a more comprehensive understanding of the nature of gun violence.

Acknowledgments

I would like to thank my mentor, Mr. Brian Lewis, for his endless help and support with my research and data analysis project on gun violence risk assessment.

References

1. *Gun violence: Prediction, prevention, and policy.* (n.d.). https://www.apa.org. https://www.apa.org/pubs/reports/gun-violence-prevention
2. Carter, P. M., Zimmerman, M. A., & Cunningham, R. M. (2021). Addressing key gaps in existing longitudinal research and establishing a pathway forward for firearm violence prevention research. *Journal of Clinical Child and Adolescent Psychology*, *50*(3), 367–384. https://doi.org/10.1080/15374416.2021.1913741
3. Writing a literature review. (n.d.). https://psychology.ucsd.edu/undergraduate-program/undergraduate-resources/academic-writing-resources/writing-research-papers/writing-lit-review.html
4. Johnson, B. T., Sisti, A. J., Bernstein, M., Chen, K., Hennessy, E. A., Acabchuk, R. L., & Matos, M. (2021a). Community-level factors and incidence of gun violence in the United States, 2014–2017. Social Science & Medicine, 280, 113969. https://doi.org/10.1016/j.socscimed.2021.113969
5. Goin, D. E., Rudolph, K. E., & Ahern, J. (2018a). Predictors of firearm violence in urban communities: A machine-learning approach. Health & Place, 51, 61–67. https://doi.org/10.1016/j.healthplace.2018.02.013
6. Ramallo, S., Camacho, M., Marín, M. R., & Porfiri, M. (2023a). A dynamic factor model to predict homicides with firearms in the United States. Journal of Criminal Justice, 86, 102051. https://doi.org/10.1016/j.jcrimjus.2023.102051
7. Data.gov home. (n.d.). Data.gov. https://data.gov/
8. Datamade. (n.d.). GitHub - datamade/census: A Python wrapper for the US Census API. GitHub. https://github.com/datamade/census
9. Population. (n.d.). https://data.ers.usda.gov/reports.aspx?ID=17827
10. Education. (n.d.). https://data.ers.usda.gov/reports.aspx?ID=17829
11. Unemployment. (n.d.). https://data.ers.usda.gov/reports.aspx?ID=17828
12. Number of mass shootings in the U.S. 1982-2023 | Statista. (2023, September 5). Statista. https://www.statista.com/statistics/811487/number-of-mass-shootings-in-the-us/
13. USDA ERS - county-level data sets. (n.d.). https://www.ers.usda.gov/data-products/county-level-data-sets/
14. Python code for transforming lat long into FIPS codes? (n.d.). Geographic Information Systems Stack Exchange. https://gis.stackexchange.com/questions/294641/python-code-for-transforming-lat-long-into-fips-codes
15. Zach. (2021). Pandas: How to Reshape DataFrame from Wide to Long. Statology. https://www.statology.org/pandas-wide-to-long/
16. Talk, M. D. (2023, January 18). Reshaping a Pandas dataframe: Long-to-Wide and vice versa. Medium. https://towardsdatascience.com/reshaping-a-pandas-dataframe-long-to-wide-and-vice-versa-517c7f0995ad
17. US Census Bureau. (2022, December 9). 2010 - 2021 County-Level estimation details. Census.gov. https://www.census.gov/programs-surveys/saipe/technical-documentation/methodology/counties-states/county-level.html
18. US Census Bureau. (2023b, September 29). Census.gov. Census.gov. https://www.census.gov/
19. U.S. Census Population Estimates — County level | Duke University Libraries. (n.d.). https://library.duke.edu/data/sources/popest
20. US Census Bureau. (2023a, June 20). County population by characteristics: 2020-2022. Census.gov. https://www.census.gov/data/tables/time-series/demo/popest/2020s-counties-detail.html
21. US Census Bureau. (2021, December 16). Census Bureau releases small area income and poverty estimates for states, counties, and school districts. Census.gov. https://www.census.gov/newsroom/press-releases/2021/saipe.html
22. Personal Income by County and Metropolitan Area, 2021 | U.S. Bureau of Economic Analysis (BEA). (n.d.). https://www.bea.gov/news/2022/personal-income-county-and-metropolitan-area-2021
23. sklearn.model_selection.cross_validate. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html#sklearn.model_selection.cross_validate
24. Welcome to GeoPy's documentation! — GeoPy 2.4.0 documentation. (n.d.). https://geopy.readthedocs.io/en/stable/
25. Poverty. (n.d.). https://data.ers.usda.gov/reports.aspx?ID=17826
26. National Bureau of Economic Research. (2023, September 29). NBER. https://www.nber.org/
27. Federal Communications Commission. (n.d.). The United States of America. https://www.fcc.gov/

Author

Rohan Singhal is a high school senior at Centennial High School in Frisco, Texas. He has knowledge of programming languages such as Java and JavaScript and has implemented machine learning models in Python. He is interested in data analysis and machine learning and hopes to study data science/statistics in college.

DOI: 10.36838/v6i8.3