Analyzing Patterns and Risk Factors of Gun Violence in the US

Muhammad Shaaf Salman
Software Engineering
FAST NUCES
Lahore, Pakistan
l216083@lhr.nu.edu.pk

Syed Farhan Jafri
Software Engineering
FAST NUCES
Lahore, Pakistan
l216074@lhr.nu.edu.pk

Muhammad Haider Khan
Software Engineering
FAST NUCES
Lahore, Pakistan
l216067@lhr.nu.edu.pk

Abstract*—This study analyzes a comprehensive dataset of over 260,000-gun violence incidents in the United States from 2013 to 2018, employing machine learning techniques, spatial-temporal analysis, and explainable AI approaches. Through exploratory data analysis, supervised and unsupervised learning models, and geographic data visualization, we uncover patterns, identify risk factors, and develop predictive models for gun violence incidents. Our findings shed light on significant risk factors, hotspots, and underlying patterns, offering insights for policymakers and law enforcement agencies to mitigate the impact of gun violence and promote public safety.

I. Introduction

A. Background and Motivation

Gun violence is a pervasive and multifaceted issue in the United States, resulting in numerous tragic incidents and profound societal consequences. Despite ongoing efforts, a comprehensive understanding of the interplay between incident characteristics, participant profiles, and spatial-temporal dynamics remains elusive. The availability of large-scale datasets capturing detailed information on gun violence incidents presents an opportunity to leverage advanced analytical techniques and machine learning approaches.

B. Objectives and Research Questions

The primary objective of this study is to leverage a comprehensive dataset on gun violence incidents in the United States to uncover patterns, identify potential risk factors, and develop predictive models that can inform prevention strategies and policy interventions. Specifically, we aim to address the following research questions:

1. What are the significant spatial and temporal patterns associated with gun violence incidents, and how do they vary across different regions and time periods?
2. Can machine learning algorithms effectively model the relationships between incident characteristics, participant profiles, and environmental factors to predict the occurrence and severity of gun violence incidents?
3. Which features or combinations of features exhibit the strongest predictive power for different types of gun violence incidents, and how can these insights guide targeted prevention efforts?
4. How can spatial-temporal analysis and visualization techniques enhance our understanding of gun violence hotspots, trends, and underlying factors contributing to these incidents?

C. Significance of the Study

This comprehensive study holds significant importance in addressing gun violence in the United States. By leveraging a large-scale dataset and advanced analytical techniques, it aims to provide data-driven insights and actionable recommendations for policymakers, law enforcement, and community organizations. The identification of spatial-temporal patterns, predictive models, and potential risk factors can aid in targeted prevention strategies, resource allocation, and intervention programs. Moreover, the integration of spatial temporal analysis and explainable AI techniques can enhance understanding, foster trust, and drive evidence-based decision-making to mitigate the impact of gun violence and promote public safety.

II. Related WORk

A. Previous Studies on Gun Violence

Existing literature has explored various aspects of gun violence, including demographic patterns, policy implications, and psychological factors. However, few studies have leveraged large-scale datasets and advanced analytical techniques to uncover comprehensive patterns, risk factors, and predictive models for gun violence incidents across the United States.

B. Machine Learning Approaches in Similar Domains

Machine learning techniques have been successfully applied in domains such as crime analysis, public health, and risk assessment, demonstrating their potential in uncovering patterns, developing predictive models, and informing data-driven decision-making processes.

III. Dataset Description

A. Data Source and Collection Method

The comprehensive gun violence dataset used in this study was obtained from the Gun Violence Archive through Kaggle. This dataset is a comprehensive record of over 260,000 US gun violence incidents recorded between January 2013 and March 2018. This dataset contains detailed information about each incident, available in CSV format. It contains date of incident, location (state, city, address), number of people killed/injured and more.

B. Dataset Characteristics and Features

The dataset contains detailed information for each gun violence incident, including the date, location (state, city/county, address), number of people killed and injured, incident characteristics (e.g., mass shooting, domestic violence, drive-by shooting), participant details (age, gender), weapon information, and other relevant notes, providing a rich set of features for analysis.

```
incident_id                  int64
date                        object
state                       object
city_or_county              object
address                     object
n_killed                     int64
n_injured                    int64
incident_url                object
source_url                  object
incident_url_fields_missing   bool
congressional_district       float64
gun_stolen                  object
gun_type                    object
incident_characteristics    object
latitude                    float64
location_description        object
longitude                   float64
n_guns_involved             float64
notes                       object
participant_age             object
participant_age_group       object
participant_gender          object
participant_name            object
participant_relationship    object
participant_status          object
participant_type            object
sources                     object
state_house_district        float64
state_senate_district       float64
dtype: object
```

C. Data Preprocessing and Cleaning

Extensive data preprocessing and cleaning steps were performed to handle missing values, remove duplicates, and ensure data consistency. This involved techniques such as imputation, dropping rows or columns with excessive missing data, encoding categorical variables, and standardizing numerical features to prepare the dataset for further analysis and modelling.

IV. Methodology

A. Exploratory Data Analysis

1) Data Visualization: Exploratory data analysis played a crucial role in gaining initial insights into the gun violence dataset. Through various visualizations, we examined the distributions of different features, identified potential outliers, and and explored spatial and temporal patterns. For instance, histograms and KDE plots helped reveal the spread and skewness of key numerical variables such as the number of individuals killed or injured, while boxplots highlighted potential outliers. Similarly, bar plots and word clouds were utilized for categorical variables like states, gun types, and incident characteristics, showcasing the most frequent occurrences in the dataset.

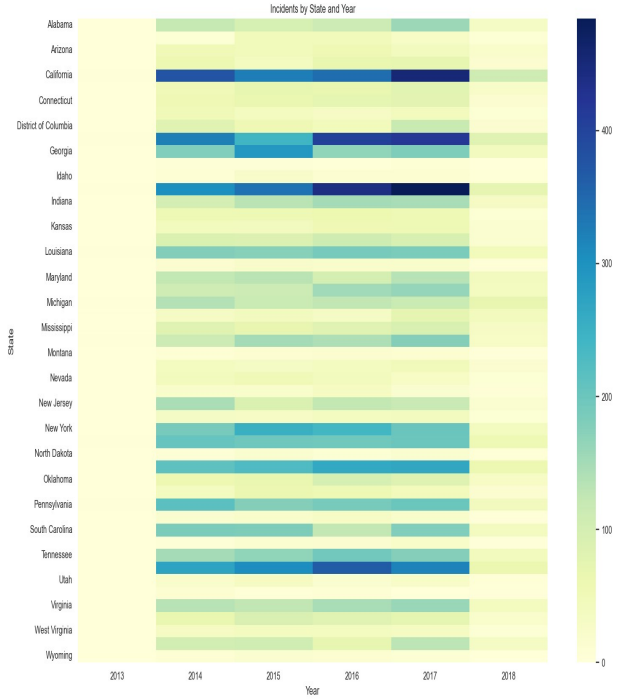

Fig 1. Incidents by State and Year

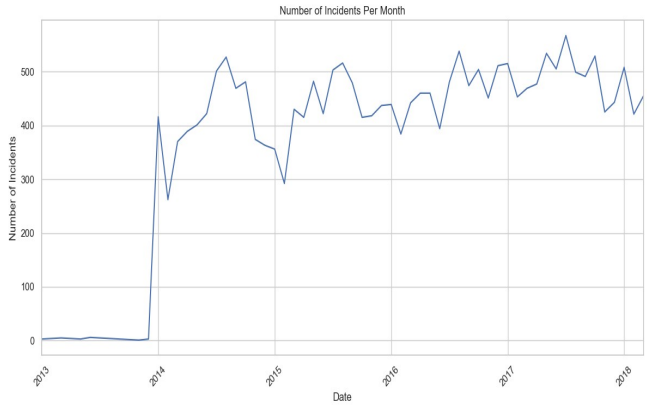

Fig 2. Number of Incidents Per Month

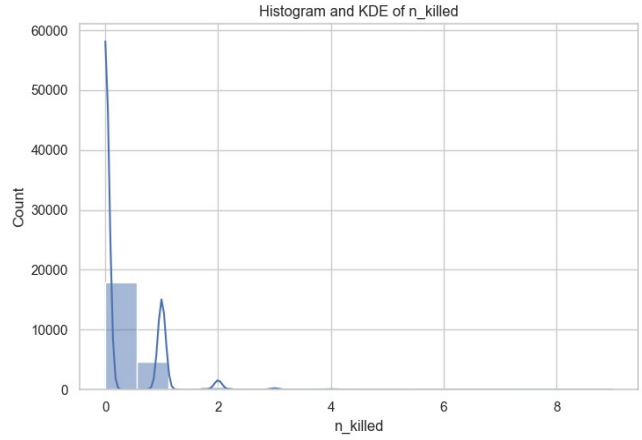

Fig 3. Number of Killed in Incidents

2) Distributions and Correlations: We examined the distributions of numerical features, such as the number of casualties and the age of participants, to identify potential skewness or outliers. Additionally, we explored correlations between various features to uncover potential relationships and dependencies that could inform feature selection and model development.

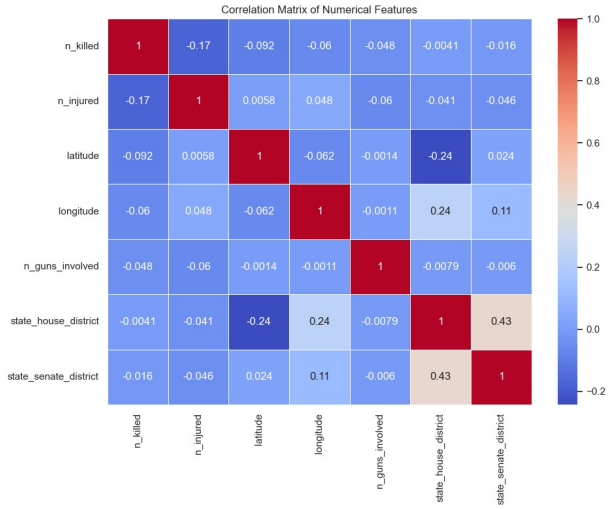

Fig 4. Correlation Matrix

3) *Spatial / Temporal Patterns*: Spatial and temporal patterns were investigated through geographic visualizations and time-series analysis. We identified potential hotspots and clusters of gun violence incidents, as well as examining seasonal or cyclical trends over time. These insights guided our subsequent spatial and temporal modelling efforts.

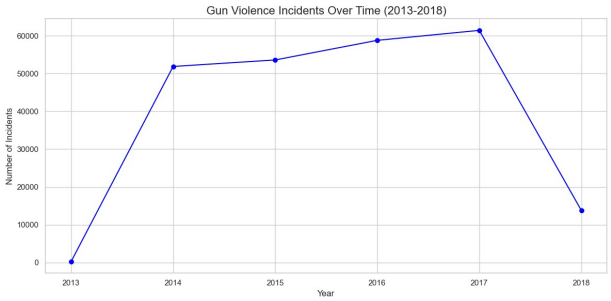

Fig 5. Temporal Pattern of Incidents

B. Feature Engineering and Selection

1) Feature Extraction and Creation: Effective feature engineering played a crucial role in our analysis. We extracted and created new features based on domain knowledge and insights gained from exploratory data analysis. This included deriving features from text descriptions, encoding categorical variables, and combining existing features to capture more complex relationships and patterns within the data.

```
Index(['n_injured', 'congressional_district', 'latitude', 'longitude',
       'n_guns_involved', 'state_house_district', 'state_senate_district',
       'n_killed'],
      dtype='object')
```

Fig 6. Feature Extraction

2) Dimensionality Reduction Techniques: To address the high-dimensional nature of the dataset and mitigate the curse of dimensionality, we employed various dimensionality reduction techniques. Principal Component Analysis (PCA) and feature selection methods, such as recursive feature elimination and correlation-based techniques, were utilized to identify the most informative features while reducing redundancy.

	PCA1	PCA2	PCA3	PCA4	PCA5
0	-0.902579	2.027094	3.400004	-1.368897	0.704606
1	-1.409030	1.907888	-2.133394	-0.880813	-0.389832
2	-0.436158	1.664019	-0.606486	-0.733160	-0.274480
3	-0.009318	0.601023	0.374455	-0.361374	0.771345
4	-0.680675	0.574758	1.888767	1.116861	0.540303
...	...	...	...	...	...
1358	-0.568681	-1.926161	0.225271	0.112712	-0.049799
1359	2.750651	0.287678	-0.304358	0.198272	0.125859
1360	-0.568681	-1.926161	0.225271	0.112712	-0.049799
1361	2.750651	0.287678	-0.304358	0.198272	0.125859
1362	2.750651	0.287678	-0.304358	0.198272	0.125859

1363 rows × 5 columns

Fig 7. PCA

C. Supervised Learning Model

1) Regression Model: To predict the number of individuals killed (n_killed) in gun violence incidents, a regression-based machine learning approach was employed. A Linear Regression model was trained on 80% of the data and tested on the remaining 20%.

```
Shape of training set (X_train): (1090, 28)
Shape of testing set (X_test): (273, 28)
Shape of training set (y_train): (1090,)
Shape of testing set (y_test): (273,)
```

Fig 8. Testing and Training Dataset

```
Shape of numeric training data: (1090, 8)

Model Coefficients: [-3.74231804e-07  2.92757482e-01 -2.64986455e-02 -3.95911533e-02
  8.36837114e-03 -1.17698124e+00 -3.62270415e-03 -1.30574585e-02]

Model Intercept: 6.140504508272253
```

Fig 9. Linear Regression Approach

D. Model Evaluation

1) Performance Metrics: Performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) were used to evaluate the model's accuracy.

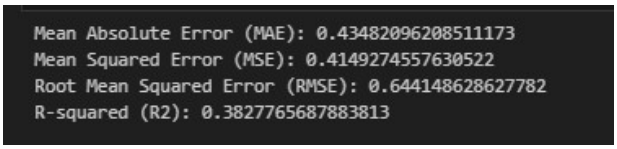

Fig 10. Performance Metrics

2) Refining: A Ridge Regression model was tuned using GridSearchCV to optimize the regularization parameter (alpha), resulting in a further improved fit.

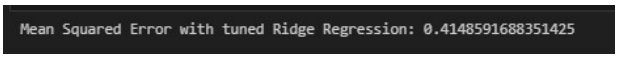

Fig 11. Tuned Mean Squared Error

E. Limitations and Future Work

While our study provides valuable insights, it is essential to acknowledge its limitations. These include potential biases in data collection, the dynamic nature of gun violence incidents, and the complexity of socioeconomic and cultural factors influencing this issue. Future work could incorporate additional data sources, explore advanced deep learning techniques, and integrate qualitative analyses to provide a more comprehensive understanding.

F. Summary of Key Findings

Our study revealed significant spatial and temporal patterns in gun violence incidents, with distinct hotspots and cyclical trends identified. Machine learning models demonstrated promising performance in predicting incident occurrence and severity, with feature importance analysis highlighting the most influential factors.

G. Implications and Recommendations

The findings from this study have important implications for policymakers, law enforcement agencies, and community organizations. Our predictive models can inform targeted prevention strategies, resource allocation, and intervention programs. Additionally, the identified risk factors and associations can guide the development of tailored educational campaigns and support services to mitigate gun violence.

H. Concluding Remarks

In conclusion, this comprehensive study leveraged the power of machine learning and advanced analytical techniques to uncover patterns, identify risk factors, and develop predictive models for gun violence incidents in the United States. By integrating explainable AI and interpretability techniques, we have fostered transparency and provided actionable insights to stakeholders. While challenges remain, our research contributes to the ongoing efforts in promoting public safety and addressing the complex issue of gun violence.

REFERENCES

Github Repository: https://github.com/farhanj21/USA-GunViolenceAnalysis
Dataset Link: https://www.kaggle.com/datasets/jameslko/gun-violence-data
Gun Violence Archive: https://www.gunviolencearchive.org/