

**ALY 6015 – INTERMEDIATE ANALYTICS**

**CHICAGO POLICE DEPARTMENT: VIOLENCE REDUCTION - SHOTSPOTTER ALERTS**

**Capstone Group 4:** Devika, Jayakumar, Sagar

**Instructor**: Zhi (Richard) He

**28th March 2024**

**Table of Contents**

Introduction3

Methods 4

Analysis5

Conclusion 20

References22

# INTRODUCTION

The ShotSpotter alerts dataset encapsulates a multifaceted tapestry of categorical and continuous variables, presenting researchers with a rich reservoir of information that demands a strategic and nuanced approach to analysis. With its extensive compilation of 20 features, comprising a minimum of 12 usable ones, and a vast dataset boasting over 3 million observations and two primary variables of interest, navigating this data landscape requires careful consideration and methodological precision. Such complexity underscores the necessity for researchers to craft research inquiries thoughtfully, ensuring alignment with the analytical capabilities of statistical models and methodologies.

One of the paramount inquiries emanating from this dataset revolves around discerning whether specific community areas exhibit statistically significant disparities in the frequency of gunshot incidents compared to others. By exploring the spatial dynamics inherent in these incidents, researchers aim to uncover potential localized hotspots of heightened activity juxtaposed with areas relatively insulated from such occurrences. This spatial analysis sheds light on geographical disparities and offers valuable insights into community safety dynamics, guiding the formulation of targeted intervention strategies and resource allocation efforts within law enforcement frameworks.

Furthermore, an additional research avenue delves into unraveling the intricate correlation between temporal aspects—such as the time of day or week—and the likelihood of gunshot incidents. By delving into temporal patterns and fluctuations, researchers endeavor to glean insights that inform resource allocation strategies and scheduling optimizations within law enforcement frameworks. Understanding when and where such incidents are more likely to occur empowers law enforcement agencies to deploy resources effectively, bolstering proactive measures for crime prevention and community safety enhancement.

RESEARCH QUESTIONS AND METHODS

1. **Community Safety Disparities:** **Are specific neighborhoods experiencing significantly higher gunshot incident rates than others?**

This question analyzes spatial dynamics to identify areas burdened by disproportionate incidents, revealing potential hotspots and regions with lower occurrences.

**Method to answer this question: EDA: We were able to illustrate the location-specific graphical plot in the exploratory data analysis by** utilizing location features.

1. **Temporal Patterns and Law Enforcement Resource Allocation: Is there a noticeable correlation between the time of day or week and the occurrence of gunshot incidents?**

This investigation seeks to uncover temporal trends, highlighting peak activity periods. Insights gained will assist in optimizing resource allocation and scheduling within law enforcement frameworks.

**Method to answer this question - EDA:** We can answer this question by performing exploratory data analysis.

1. **Predictive Model for Incident Discrimination: Can we develop a robust predictive model capable of distinguishing between incident types, such as single and Multiple gunshots, using factors like location and time?**

This question aims to harness predictive analytics to create a sophisticated model that facilitates swift and targeted responses, enhancing community safety and law enforcement strategies.

**Method to answer this question: Predictive Models:** We have tried predictive models such as GLM—logistic Regression, linear regression, and Decision-Tree Models.

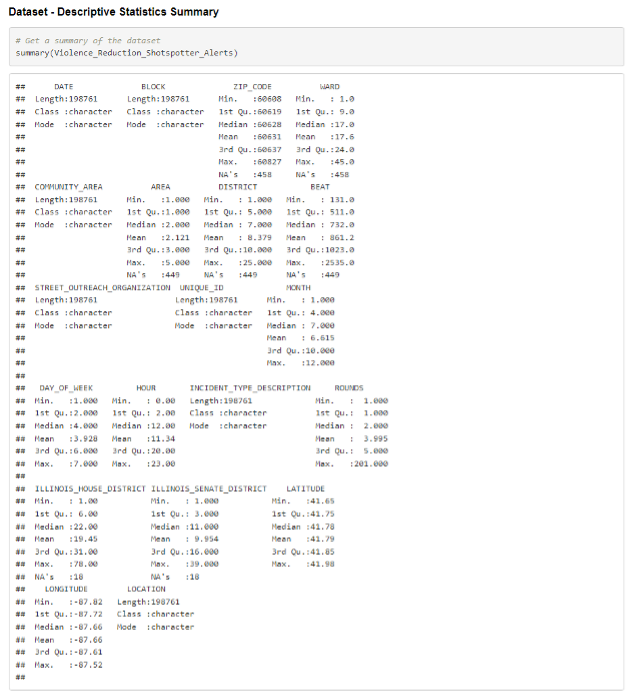
Regarding our current progress, we have addressed **Question 2** in our initial analysis, leveraging predictive modeling methods. However, **Question 1** necessitates further investigation involving geographical mapping to pinpoint the most affected areas accurately. This process requires additional time as the initial heatmap analysis did not provide conclusive insights. **Question 3** was addressed comprehensively in the final report, employing advanced predictive modeling techniques to refine our understanding of incident discrimination.

# EXPLORATORY DATA ANALYSIS

The dataset contains around 200,000 records, detailed across 20 fields that capture a range of information, from temporal aspects like dates to locational data such as blocks, ZIP codes, and community areas, along with specifics of the incidents like the type and number of rounds fired, as well as precise geographical markers given by latitude and longitude. This extensive data collection affords an in-depth exploration of firearm discharge events over a specific timeframe.

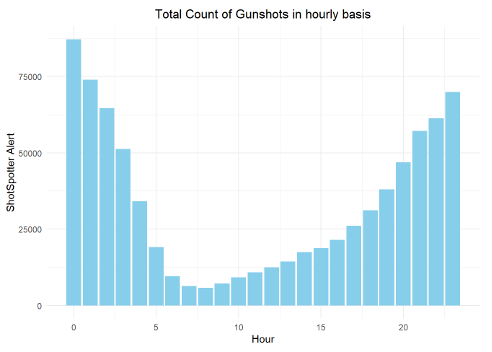
## INTERPRETATIONS OF OUTPUTS AND RESULTS

**Descriptive Statistics:** Initial data assessments will furnish a broad picture of how gunshot incidents are dispersed in terms of timing, geographical placement, and the nature of the occurrences (like the amount of ammunition expended). For example, discovering a higher rate of incidents within particular ZIP codes or neighborhoods could point to concentrated zones of gun-related activities. These insights would be instrumental in devising focused intervention strategies.

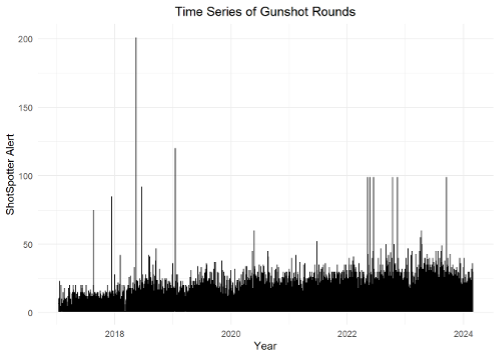


**Question - Is there a noticeable correlation between the time of day or week and the occurrence of gunshot incidents?**

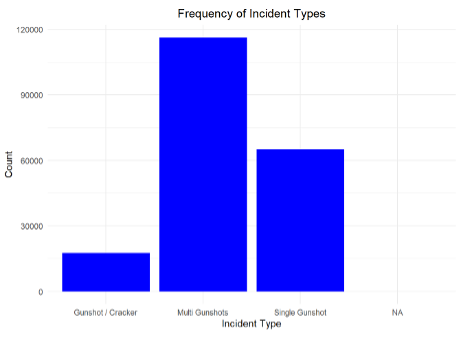
Temporal Trends: Examining the dataset over various time intervals could reveal cyclic or periodic trends, such as a surge in incidents in specific months or at particular times within a day. For instance, a consistent rise in gunshot alerts at night could provide valuable insights for law enforcement agencies and community groups, indicating potential times for heightened patrol presence or intensified community outreach programs.



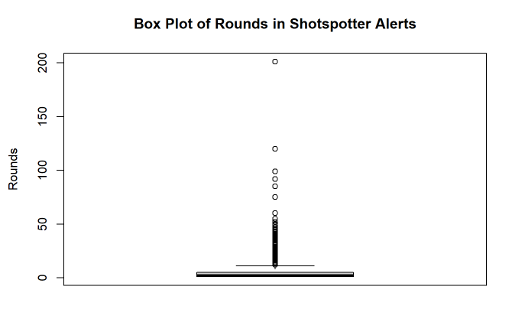
The graph illustrates a trend where ShotSpotter alerts predominantly occur at night, with a noticeable escalation coinciding with the evening hours. This uptick may primarily be attributed to the detonation of firecrackers, which are likely to activate the ShotSpotter sensors.



**Analysis of Incident Categories and Ammunition Discharged:** Delving into the specifics of incident classifications, such as single versus multiple gunshots, alongside tallying the ammunition expended, sheds light on the severity and character of gun-related violence within distinct locales. Regions reporting a higher frequency of numerous gunshot occurrences may be grappling with more intense violent situations, which could call for more concentrated and strategic intervention measures.

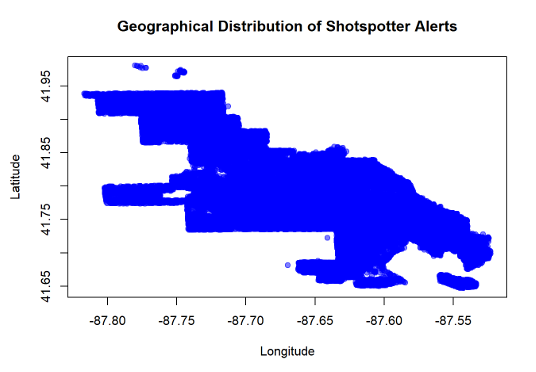


**Outlier Analysis:** The variable “Rounds,” indicating the count of ShotSpotter alerts logged by the Chicago Police Department, shows several instances extending beyond 25 alerts. Such high frequencies of alerts, ranging between 25 to 200, are atypical for gunshots, suggesting these may be anomalies. These outliers are likely attributable to sources other than gunfire, such as firecrackers.



**Question - Are specific neighborhoods experiencing significantly higher gunshot incident rates than others?**

**Geospatial Analysis:** The dataset's geospatial information enables the identification of clusters with prevalent gun-related incidents. Localities recording elevated incidents may be targeted as focal points for violence mitigation initiatives, potentially involving enhancing surveillance measures or amplifying community engagement efforts.



# MODELS

**RATIONALE:**

The Chicago Police Department: Violence Reduction—Shotspotter Alerts dataset is considered one of the best datasets we have worked on. Its information on location, occurrences, time, etc., is obvious. As explained earlier, the dataset is maintained bi-weekly. As the dataset explicitly expresses the response variables, “INCIDENT\_TYPE\_DESCRIPTION” has been chosen as the response variable.

The following snapshot illustrates the head of the response variables. Since the “Rounds” feature is a numeric data type with multiple outliers, we have converted the “Incident type description” categorical variable into a binomial variable as Single Gunshot = 0 and Multiple Gunshot and Firecracker = 1.

**Response Variable for GLM Logistic Regression and Decision-Tree Classification Models:**

**A screenshot of a computer

Description automatically generated**

**Response Variable for Linear Regression Model:**

**A screenshot of a phone

Description automatically generated**

**Predictor Variables:**

* ZIP\_CODE
* MONTH
* DAY\_OF\_WEEK
* ROUNDS
* LATITUDE / LONGITUDE

**CHOSEN MODELS**

**Classification Model:**

We have performed the following models under classification for our newly added feature “Incident\_type\_code, " a binomial variable.

* GLM – Logistic Regression
* Decision-Tree

**Regression Model:**

We have performed the following regression model to understand the model’s performance with predictions using the numerical response variable “ROUNDS.”

* Linear Regression

## MODEL ANALYSIS

**GLM-Logistic Regression Model**

1. **Significant Predictors:**
   * The variables MONTH, DAY\_OF\_WEEK, ROUNDS, and LONGITUDE are statistically significant predictors of the incident type, as indicated by their low p-values (p < 0.05).

**A screenshot of a computer

Description automatically generated**

* + Specifically, the coefficients for MONTH, DAY\_OF\_WEEK, and LONGITUDE are positive, suggesting that higher values of these variables are associated with increased odds of a particular incident type.

1. **Variable Importance:**
   * The variable ROUNDS is highly significant, with a large coefficient estimate of approximately 4.63. This indicates that the number of rounds (possibly referring to some measure of police activity or patrol) substantially impacts the incident type.
   * Additionally, the coefficient for LONGITUDE is significant, suggesting that geographical location (longitude) plays a role in determining the incident type.
2. **Model Fit:**
   * The model's goodness-of-fit statistics, such as the AIC (Akaike Information Criterion) and residual deviance, suggest that the model fits the data reasonably well. The AIC value of 53475 indicates that this model provides a relatively good balance between fit and complexity.
   * The residual deviance, which compares the model's fit with a null model (no predictors), is considerably lower than the null deviance, indicating that the model explains a significant portion of the variability in the response variable.

The logistic regression model suggests that MONTH, DAY\_OF\_WEEK, ROUNDS, and LONGITUDE are essential factors in predicting the incident type. The model provides valuable insights into the relationships between these predictors and the outcome variable, contributing to our understanding of the factors influencing incident types.

**Confusion Matrix– GLM-Logistic Regression:**

**A white squares with black text

Description automatically generated**

* The logistic regression model accurately predicted 36,563 instances as positive and 19,185 instances as negative.
* However, it misclassified 125 instances as positive when they were negative (false positives).
* Additionally, 3,491 instances were incorrectly classified as negative when they were positive (false negatives).
* Overall, the model exhibits strong performance with high true positive and true negative rates but has room for improvement in minimizing false positives and false negatives.

**ROC Curve – GLM-Logistic Regression:**

* The ROC curve depicts the sensitivity (actual positive rate) versus 1-specificity (false positive rate) across different threshold settings.

A graph of a curve

Description automatically generated

* It illustrates the model's ability to distinguish between positive and negative instances.
* A high AUC value of approximately 0.925 for the logistic regression model indicates robust discriminatory power, highlighting its effectiveness in differentiating various incident types.

**Decision-Tree Model**

* The decision tree model's structure is presented, highlighting the variables used for splitting and the resulting terminal nodes.
* An AUC value derived from ROC curve analysis showcases the model's discriminative prowess, with an AUC of around 0.995, indicating exceptional predictive capability.
* The confusion matrix delineates the model's classification performance, detailing the counts of true positives, true negatives, false positives, and false negatives. It demonstrates high accuracy with minimal misclassifications.

**Confusion Matrix – Decision-Tree:**

**A white rectangular box with black text

Description automatically generated**

* True Positives (TP): 36,563 instances correctly classified as positive.
* True Negatives (TN): 19,185 instances correctly classified as negative.
* False Positives (FP): 125 instances incorrectly classified as positive.
* False Negatives (FN): 3,491 instances incorrectly classified as negative.

The model achieves high accuracy in positive and negative classifications, suggesting robust performance.

**ROC Curve – Decision-Tree:**

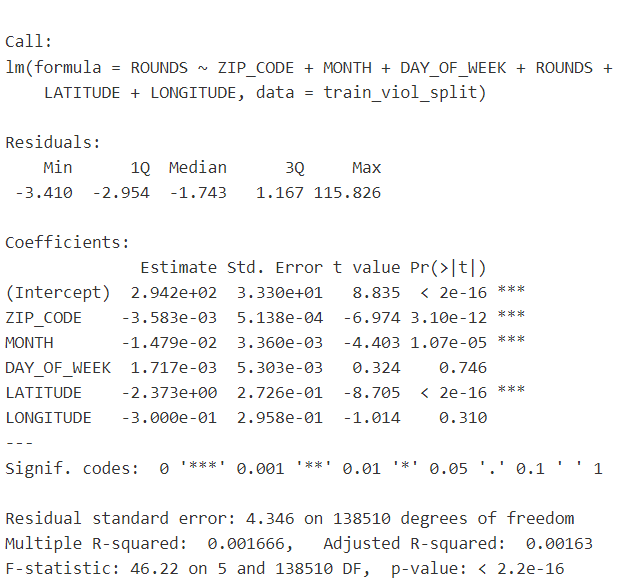
* The ROC curve provides insights into the sensitivity-specificity trade-off across various threshold values.
* The decision tree model's outstanding predictive performance is highlighted by an AUC of approximately 0.995, surpassing other regression models in discriminatory power.

A diagram of a test set

Description automatically generated

**Linear Regression**

**Model Summary:**

****

* **Model Coefficients:**
  + The linear regression model indicates that ZIP\_CODE, MONTH, LATITUDE, and LONGITUDE have statistically significant coefficients (p < 0.05), suggesting they are essential in predicting the number of rounds.
  + For every one-unit increase in ZIP\_CODE, the number of rounds decreases by approximately 0.0036 units.
  + An increase of one unit in MONTH decreases around 0.0148 units in the number of rounds.
  + LATITUDE has a coefficient of approximately -2.373, indicating that an increase in LATITUDE is associated with a decrease in the number of rounds.
  + LONGITUDE, however, does not significantly affect the number of rounds (p = 0.310).

A screenshot of a data

Description automatically generated

* **Model Performance:**
  + The model's performance metrics:
    - Residual Standard Error (RSE): 4.346
    - Mean Squared Error (MSE): 11.3201
    - Root Mean Squared Error (RMSE): 3.364536
    - Mean Absolute Error (MAE): 3.327261
    - Multiple R-squared: 0.001666
  + The RSE reflects the average deviation of the observed values from the predicted values, providing a measure of the model's accuracy.
  + The low R-squared value (0.001666) suggests that the model explains only a tiny proportion of the variance in the number of rounds.
* **Comparison between Trained and Test Data:**
  + The comparison table highlights differences in model performance between trained and test data:
    - Trained Data:
      * Residual Standard Error (RSE): 0.4117
      * Multiple R-squared: 0.2291
      * Adjusted R-squared: 0.2291
    - Test Data:
      * Mean Squared Error (MSE): 0.1695297
      * Root Mean Squared Error (RMSE): 0.4117399
      * Mean Absolute Error (MAE): 0.3795446

The results underscore its coefficients, performance metrics, and the disparities observed between trained and test data. Although some predictors show significance, the model's explanatory power is limited, as indicated by the low R-squared value.

**Model Comparison:**

|  |  |
| --- | --- |
| Model | Key Findings |
| GLM-Logistic Regression | - Significant predictors: MONTH, DAY\_OF\_WEEK, ROUNDS, and LONGITUDE have low p-values, indicating their significance. |
| - ROUNDS and LONGITUDE show high importance. |
| - The model exhibits a good fit, as indicated by the AIC and residual deviance. |
| - The confusion matrix shows strong performance, with high true positive and true negative rates, but room for improvement in minimizing false positives and false negatives. |
| - ROC curve demonstrates robust discriminatory power with an AUC of approximately 0.925. |
| Decision-Tree | - A decision tree structure was presented, showcasing variable importance and terminal nodes. |
| - A high AUC of around 0.995 indicates exceptional predictive capability. |
| - Confusion matrix highlights high accuracy with minimal misclassifications. |
| - ROC curve underscores outstanding predictive performance. |
| Linear Regression | - Significant coefficients for ZIP\_CODE, MONTH, and LATITUDE. |
| - The model's performance metrics indicate limited explanatory power with a low R-squared value. |
| - Comparison between trained and test data reveals differences in model performance. |

CONCLUSION

Our detailed analysis explores three models to understand firearm incidents detected by ShotSpotter alerts. The GLM-Logistic Regression model identifies key predictors like MONTH, DAY\_OF\_WEEK, ROUNDS, and LONGITUDE, emphasizing their role in predicting incident types. While this model shows vital performance metrics and discriminatory power, there's potential for improvement in reducing false positives and negatives. In contrast, the Decision-Tree model excels in predictive capability, boasting high accuracy and minimal misclassifications. However, the linear regression model, which highlights significant predictors, falls short of explaining variations in incident occurrences, indicating the need for further refinement. These insights underscore the importance of selecting the right modeling approach tailored to address specific business inquiries effectively.

# REFERENCES

* Data.gov, 2024, <https://catalog.data.gov/dataset/violence-reduction-shotspotter-alerts>