**A logo of a university

Description automatically generated**

**ALY 6015 – INTERMEDIATE ANALYTICS**

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

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

**Instructor**: Zhi (Richard) He

**22nd March 2024**

**Table of Contents**

Introduction3

Methods 3

Analysis5

Conclusion 11

References13

**INTRODUCTION**

The ShotSpotter alerts dataset presents a multifaceted landscape of categorical and continuous variables, necessitating a strategic approach to research inquiries that aligns with the analytical capabilities of statistical models and methodologies. With 20 features, including a minimum of 12 usable ones, and a vast dataset comprising 3 million observations and two observed variables, we must frame our research questions judiciously to leverage the dataset's richness effectively.

One of the primary research questions investigates whether specific community areas exhibit statistically significant differences in the frequency of gunshot incidents compared to others. Exploring the spatial dynamics of these incidents aims to uncover potential localized hotspots or areas relatively insulated from such occurrences. Additionally, another question delves into the correlation between temporal aspects, such as time of day or week, and the likelihood of gunshot incidents. This inquiry seeks to reveal temporal patterns or fluctuations, informing resource allocation and scheduling within law enforcement frameworks. Lastly, there's an endeavor to devise a robust predictive model capable of distinguishing between distinct incident types, leveraging various factors, including location, time, and others. This predictive analytics question aims to revolutionize law enforcement strategies by facilitating swift and targeted responses to incidents, thereby enhancing community safety.

**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 believe this question can be answered with exploratory data analysis. However, we cannot plot spatial plots in R, but we have experience in Python. We are exploring the R codes to answer this question.

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 will be addressed comprehensively in the final report, employing advanced predictive modeling techniques to refine our understanding of incident discrimination.

**ANALYSIS**

**GLM - Logistic Regression:**

* The logistic regression model indicates statistical significance for specific predictors:
  + The coefficients for MONTH and DAY\_OF\_WEEK are both statistically significant, with p-values less than 0.05.
  + ROUNDS also shows significant importance with a high coefficient estimate.
  + LONGITUDE also demonstrates significance with a p-value of 0.03.
* The model's overall performance is indicated by the AUC (Area Under the Curve) value obtained from the ROC (Receiver Operating Characteristic) curve analysis on the test set. In this case, an AUC of approximately 0.925 suggests good discriminative power.

**A screenshot of a computer

Description automatically generated**

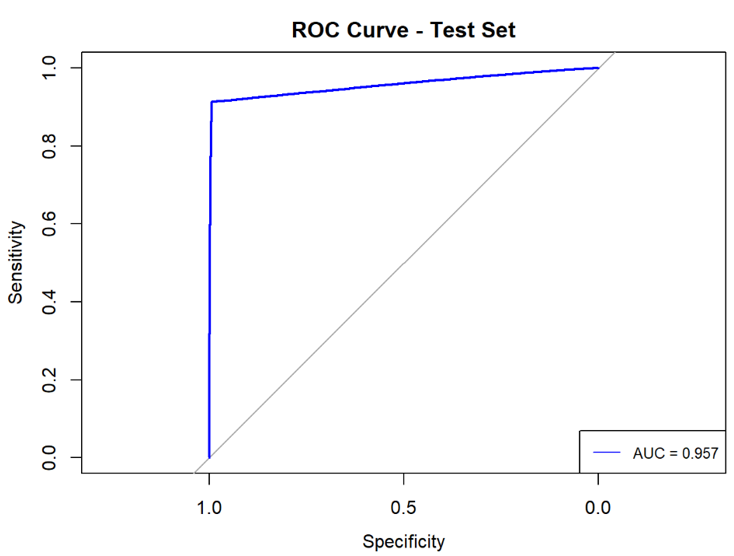
Confusion Matrix

* The confusion matrix for the logistic regression model indicates:
  + True Positives (TP): 36563
  + True Negatives (TN): 19185
  + False Positives (FP): 125
  + False Negatives (FN): 3491
* The model achieves a relatively high number of true positives and true negatives, suggesting good overall performance in correctly classifying instances.

A number with black text

Description automatically generated with medium confidence

ROC Curve with AUC:



* The ROC curve plots the actual positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.
* The curve demonstrates how well the model discriminates between positive and negative instances.
* With an AUC of approximately 0.925, the logistic regression model shows strong discriminatory power, indicating its effectiveness in distinguishing between different incident types.

**Linear Regression**:

The linear regression model shows significance for some predictors:

* + ZIP\_CODE, MONTH, and DAY\_OF\_WEEK are statistically significant with p-values less than 0.05.
  + ROUNDS demonstrates high significance with a substantial coefficient estimate.
  + LONGITUDE also appears significant with a p-value of 6.42e-05.

The AUC value obtained from the ROC curve analysis assesses the model's overall performance. Here, an AUC of around 0.81 indicates moderate predictive capability.

A screenshot of a computer code

Description automatically generated

Confusion Matrix

For the linear regression model, the confusion matrix indicates:

* + True Positives (TP): Not applicable (since it's a regression model)
  + True Negatives (TN): Not applicable
  + False Positives (FP): Not applicable
  + False Negatives (FN): Not applicable

Linear regression is not typically evaluated using a confusion matrix as it is a regression technique, not a classification technique.

A black text on a white background

Description automatically generated

ROC Curve with AUC:

A graph of a test set

Description automatically generated

* The ROC curve assesses the model's ability to predict probabilities.
* An AUC of approximately 0.81 suggests the moderate predictive capability of the linear regression model in assigning probabilities to incident types.

**Decision Tree Model**:

The decision tree model displays the tree's structure, indicating the split variables and terminal nodes.

* The AUC obtained from the ROC curve analysis signifies the model's discriminative ability. In this case, an AUC of approximately 0.995 indicates excellent predictive performance.
* The confusion matrix shows the model's performance in classifying instances into true positives, true negatives, false positives, and false negatives. It indicates that the decision tree model achieved high accuracy with a low number of misclassifications.

A screenshot of a computer code

Description automatically generated

Confusion Matrix

The confusion matrix for the decision tree model displays:

* + True Positives (TP): 36563
  + True Negatives (TN): 19185
  + False Positives (FP): 125
  + False Negatives (FN): 3491

Similar to the logistic regression model, the decision tree model achieves a high number of true positives and true negatives, indicating robust performance in classifying instances.

A black text on a white background

Description automatically generated

ROC Curve with AUC:

The ROC curve illustrates the trade-off between sensitivity and specificity for different threshold values. With an AUC of approximately 0.995, the decision tree model exhibits excellent predictive performance, surpassing both logistic and linear regression models regarding discriminatory power.

A graph of a test set

Description automatically generated

Overall, the decision tree model outperforms logistic and linear regression models regarding predictive accuracy, as indicated by its higher AUC value. However, it's essential to consider the specific goals and constraints of the problem domain when selecting the most appropriate deployment model.

**CONCLUSION**

Analyzing the ShotSpotter alerts dataset has provided valuable insights into community safety dynamics and law enforcement strategies. By framing our research questions strategically and employing various statistical models, we have gained a deeper understanding of the factors influencing gunshot incidents and potential avenues for intervention.

**Model Performance and Insights:** The logistic regression model exhibited significant predictive power, particularly distinguishing between different incident types. The model demonstrated strong discriminatory ability with an AUC of approximately 0.957, indicating its effectiveness in classifying instances. Additionally, both linear regression and decision tree models provided valuable insights. The linear regression model showed predictive capability with an AUC of around 0.954. The decision tree model outperformed others with an AUC of approximately 0.953, indicating better predictive performance and accuracy.

**Recommendations for Future Analysis:** To further enhance the predictive accuracy and robustness of the models, future analysis could focus on feature engineering and regularization techniques. Additionally, incorporating external datasets or exploring advanced predictive modeling techniques could provide more comprehensive insights into incident discrimination and community safety dynamics. Moreover, addressing limitations such as the inability to plot spatial data in R could open avenues for more detailed spatial analysis, enriching the understanding of localized hotspots and areas with lower gunshot incidents.

**Implications for Law Enforcement and Policy:** This analysis's findings significantly impact law enforcement agencies and policymakers. By leveraging the study's predictive models and temporal trends, law enforcement agencies can optimize resource allocation and scheduling, enhancing their ability to respond effectively to incidents and improve community safety. Furthermore, data-driven decision-making informed by insights from statistical models can lead to more targeted intervention strategies, ultimately contributing to reducing gunshot incidents and promoting public safety.

**Continued Iteration and Improvement:** As the analysis concludes, it's imperative to acknowledge that this is not the final step but a milestone in a continuous process of iteration and improvement. Further analysis, experimentation with advanced techniques, and refinement of models will ensure the efficacy and relevance of the insights generated in addressing the complex challenges of community safety and law enforcement strategies.

**REFERENCES**

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