Community Safety

DS401 Project Website

**Team Members**: Neha Maddali, Alexis Maldonado, Pramit Vyas

**Client**: Shawn F Dorius, PhD

**Course Name**: DS401

**Instructor**: Dr. Adisak

# Project Contacts

Shawn F Dorius: [sdorius@iastate.edu](mailto:sdorius@iastate.edu)

Kelsey R Van Selous: [kvansel@iastate.edu](mailto:kvansel@iastate.edu)

Matthew J Voss: [mjvoss@iastate.edu](mailto:mjvoss@iastate.edu)

# Project Team

Neha Maddali: [nmaddali@iastate.edu](mailto:nmaddali@iastate.edu)

Alexis Maldonado: [ahm@iastate.edu](mailto:ahm@iastate.edu)

Pramit Vyas: [pvvyas@iastate.edu](mailto:pvvyas@iastate.edu)

# Stakeholders

Public Science Collaborative

# Project Description

## Introduction

In recent years, communities nationwide have dealt with alarming increases in both property and violent crimes. Moreover, the persistent scourges of substance abuse and firearm-related fatalities have cast a shadow over the health and safety of residents. Recognizing these pressing issues, there arises a crucial need for a sophisticated local data surveillance system equipped with predictive capabilities. Such a system would not only inform decision-making processes but also facilitate targeted interventions to safeguard community well-being.

The essence of this project lies in the development and deployment of a predictive model: Predict-Align-Prevent. This aims to fortify community health and safety initiatives. At its core, this endeavor seeks to establish comprehensive data pipelines tailored to local contexts, meticulously capturing temporal, spatial, and risk-related events. These events encompass a spectrum of critical incidents, ranging from overdoses to firearm-related incidents, and are intricately linked to the unique risk and protective factors prevalent in each community.

Collaborating closely with the Public Science Collaborative which consists of esteemed public health and safety experts, our endeavor is rooted in a commitment to empower communities with actionable insights derived from robust data analytics. Partnering with the PSC, we endeavor to construct a suite of predictive models under the banner of PREDICT. However, PREDICT merely marks the inception of our multifaceted approach.

Beyond predictive analytics, our methodology extends to the pivotal stages of ALIGN and PREVENT. Following the generation of accurate predictions, our concerted efforts converge on aligning resources and interventions with identified risk factors. Through close engagement with local decision-makers, we endeavor to orchestrate a strategic allocation of resources aimed at preemptively addressing potential threats. By aligning our interventions with community needs and dynamics, we try to catalyze proactive measures aimed at mitigating shootings, overdoses, and intoxicated driver incidents.

In essence, our project represents a holistic endeavor to harness the power of data-driven insights in fortifying the fabric of community health and safety. By fusing cutting-edge predictive analytics with community-centered interventions, we aspire to usher in a new era of resilience and well-being across neighborhoods throughout the county.

## Methodology

Below is a workflow of this project:

A diagram of a model

Description automatically generated

To start, exploratory data analysis was conducted on the call for service data. This was to understand the distribution of data and the variables presented. Analysis was done using R code with libraries like GGPlot and RShiny.

Following this, 3 models were developed to predict the hourly number of calls.

### Logistic Regression

Logistic regression serves as a statistical technique for binary classification tasks, especially in predicting outcomes with two possible categories. Its simplicity, interpretability, and effectiveness in modeling relationships between independent variables and the probability of a specific outcome make it a preferred choice.

When it comes to predicting theft or shoplifting calls, a logistic regression model becomes a significant tool for understanding temporal patterns and trends in such incidents. By using the dataset spanning from March 2018 to December 2022, trends are effectively captured, ensuring the model's accuracy in predictions.

The model uses key features like the Block Group GEOID, hour of the day, and day of the week. These features enable the model to find peak hours, daily variations, and weekly patterns in theft incidents.

The model building process follows a systematic approach:

1. Data Preparation: The dataset is split into a training and test set, with 70% of the data going to the training and 30% for testing.
2. Logistic Model: A logistic regression model with the default parameters is built, serving as a baseline. After reviewing the model summary, the unnecessary predictors were removed, resulting in a “final” model.
3. Performance Evaluation: The model's performance is evaluated using metrics such as accuracy, F1 score, and confusion matrix, highlighting its predictive power and effectiveness.
4. Evaluation: The final logistic regression model's accuracy is examined to assess its ability to predict theft calls accurately. Accuracy is important as it indicates the proportion of correct predictions made by the model.

The final logistic regression model for predicting theft or shoplifting had an accuracy of 0.68 and an F1 score of 0.81, compared to the original model's accuracy of 0.61 and F1 score of 0.75. While accuracy is informative, it might not sufficiently reflect model effectiveness in imbalanced datasets. The F1 score of 0.81 indicates a balanced performance between precision and recall, crucial for correctly identifying positive samples while minimizing misclassifications of negative ones.

### Random Forest

Random Forest is a versatile and powerful machine learning algorithm commonly used for both classification and regression tasks. It belongs to the ensemble learning family and is comprised of a collection of decision trees. Each decision tree is constructed by recursively partitioning the feature space into increasingly homogeneous subsets based on the values of input features. This process continues until a stopping criterion is met, typically defined by a maximum depth or minimum number of samples per leaf node.

But what sets Random Forests apart is its ensemble nature. Rather than relying on a single decision tree, Random Forest aggregates predictions from multiple decision trees to arrive at a final prediction. This is achieved through a process known as bagging (Bootstrap Aggregating), where each decision tree is trained on a random subset of the training data, sampled with replacement.

Moreover, Random Forest introduces an additional layer of randomness by selecting a random subset of features at each split point during the construction of each decision tree. This helps to decorrelate the individual trees and further improve the model’s performance by reducing overfitting.

The primary objective of the Random Forest Regression model is to predict the number of Nuisance calls per hour. This prediction task is essential for understanding temporal patterns and trends in category-based incidents, which can aid in resource allocation, proactive planning, and decision-making processes for public safety agencies.

Below is the process used to build an effective random forest model for the call for service data.

1. Data Preparation
   1. Split Data: The dataset was divided into training and test sets using a 70-30 split. The training set, comprising 70% of the data, was used for model training, while the remaining 30% served as the holdout test set for model evaluation.
2. Random Forest Model
   1. Built Initial Model with Default Hyperparameters: An initial Random Forest Regression model was constructed using default hyperparameters. This baseline model served as the starting point for subsequent optimization and tuning efforts.
   2. Evaluated Performance Metrics: The performance of the initial model was evaluated using various metrics, including mean absolute error (MAE), mean squared error (MSE), and R-squared. These metrics provided insights into the model’s predictive accuracy and goodness of fit.
3. Hyperparameter Tuning
   1. Tuned Model using Cross-Validation (5 folds): Hyperparameters of the Random Forest model was fine-tuned using 5-fold cross-validation. Specifically, the ‘mtry’ hyperparameter, which determines the number of variables randomly sampled as candidates at each split, was optimized to improve model performance.
   2. Selected Best ‘mtry’ Hyperparameter: The best ‘mtry’ hyperparameter value was selected based on cross-validation results, maximized the model’s predictive accuracy on the training data.
4. Tuned Model
   1. Rebuilt Model with Tuned Parameters: The Random Forest Regression model was rebuilt using the selected optimal hyperparameters obtained from the tuning process. This refined model aimed to enhance predictive performance by incorporating the optimized parameter values.
5. Bagging
   1. Applied to Tuned Random Forest Model: Bagging, a technique that aggregates multiple instances of a model to improve stability and accuracy, was applied to the tuned Random Forest model. Bagging helps reduce variance and prevent overfitting by averaging predictions from multiple trees.
6. Evaluation
   1. Calculated Out-of-Bag Error: The final bagged Random Forest model’s performance was assessed using out-of-bag (OOB) error estimation. OOB error provides an unbiased estimate of the model’s prediction error without the need for a separate validation set.

By following this systematic model building process, we aimed to develop a robust Random Forest Regression model capable of accurately predicting the number of category-specific calls per hour, thereby aiding in effective resource allocation and decision-making for public safety initiatives. The resulting model had an out of bag misclassification error rate of 11.09%.

### Neural Network

A basic Neural network is like a linear Perceptron model.

Perceptron assigns random weights, classifies samples using a sign function, and updates weights based on misclassifications until convergence or a max iteration limit.

Neural networks expand on this by connecting perceptrons in layers, using complex functions like ReLu.

Convolutional Neural Networks (CNNs) detect spatial patterns in data efficiently.

Recurrent Neural Networks (RNNs) capture sequential dependencies, suitable for time series forecasting.

Prior work includes similar projects in Pittsburgh using feed-forward neural networks for crime prediction and an article on crime hotspot forecasting using RNNs.

Below is the process to build the model.

Preprocess Data: Handle missing values, normalize numerical features, and encode categorical variables.

Design Neural Network Architecture: Choose suitable neural network type (e.g., MLP, CNN, RNN), determine layers and neurons, select activation functions and loss functions, define optimizer, and set hyperparameters.

Train the Neural Network: Feed training data, adjust weights iteratively to minimize loss, monitor performance on validation set, experiment with hyperparameters, and stop training when performance degrades.

Evaluate the Model: Assess performance on test set using appropriate metrics, analyze discrepancies, consider additional evaluation techniques.

## Final Product

Our final product is a comprehensive and insightful report detailing our analysis of police call for service data. This meticulously crafted report encompasses our exploration of temporal patterns, influential factors, and predictive modeling techniques applied to the dataset spanning from 2018 to 2023. Through the integration of machine learning algorithms such as Logistic Regression, Random Forest, and Neural Networks, we have elucidated significant predictors of incident counts and provided valuable insights into the underlying dynamics of incident reporting. Our report not only highlights key findings, such as peak call frequencies during certain months and hours, but also outlines future avenues for research and data collection to further enhance the accuracy and applicability of predictive models. With its clear presentation of results, methodological approach, and actional recommendations, our report serves as a valuable resource for stakeholders interested in leveraging police call for service data to inform decision-making and enhance community safety measures.

## Findings

Given the call service data from March 2018 to December 2023 across Mason City, Clear Lake City, and Cerro Gordo County, there are various categories of call types that our client had aggregated. These categories include Other, Nuisance, Health, Domestic, Blight, Violence, Substance-Off Premise, Substance-Driving, and Substance-On Premise.

From the data, it was noted that Mason City PD had the most amount of service call data, following Clear Lake PD and then Cerro Gordo County Sherriff. There was a consistent trend across the 5 years of data indicating that the calls increase as approaching May, where the peak of number of calls is in July and then decreases as approaching December. The same can be said about the hours of the day: the number of calls increases as approaching the afternoon, where the peak of the number of calls is at 3 pm and then decreases as approaching the end of the day. We can see this in certain areas such as around the lake and downtown Mason City as it gets later into the day.

It can also be noted that the most number of calls are during the daytime of the weekdays: Monday, Tuesday, Wednesday, Thursday, and Friday.

Between the three call sources (Mason City, Clear Lake, and Cerro Gordo), Cerro Gordo County Sherriff received the most service calls related to Substance-Driving and Substance-On Premise across the 5 years of data. This can be explained by the fact that the Sherriff Department responds to more calls that are made from the I-35 and US-65 routes where many of these Substance-Driving calls are made from.

Substance-Off Premise calls were well split among the Mason City PD and Clear Lake PD across the 5 years of data. This can be explained from our developed heat maps that clearly indicate that many of these incidents occur within the towns of Mason City and Clear Lake, hence 911 calls being made directly to their departments.

In our modeling endeavors, we employed various machine learning algorithms, including Logistic Regression, Random Forest, and Neural Networks, to predict incident counts based on a range of features within the dataset. In Logistic Regression, features such as cfs\_hour, cfs\_weekday and block\_grou\_GEOID emerged as significant predictors, offering insights into the temporal and spatial dynamics of incident reporting. Similarly, Random Forest analysis highlighted the significance of cfs\_hour as the most influential predictor, emphasizing the temporal aspect in incident occurrence. In the case of Neural Network modeling, lag predictors along with cfs\_hour and cfs\_day demonstrated considerable significance, underscoring the importance of temporal and historical patterns in predicting incident counts.