Community Safety

DS401 Final Report Document

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

**Client**: Shawn F Dorius, PhD

**Course Name**: DS401

**Instructor**: Dr. Adisak

Table of Contents

[Introduction 3](#_Toc165624394)

[Motivation and Background 3](#_Toc165624395)

[Data Overview 4](#_Toc165624396)

[Exploratory Data Analysis 6](#_Toc165624397)

[High-Level Summary of EDA 17](#_Toc165624398)

[Logistic Regression Model 18](#_Toc165624399)

[Background 18](#_Toc165624400)

[Model Specifications 18](#_Toc165624401)

[Model Building Process 19](#_Toc165624402)

[Model Results 20](#_Toc165624403)

[Random Forest Model 20](#_Toc165624404)

[Background 20](#_Toc165624405)

[Model Specifications 21](#_Toc165624406)

[Model Building Process 22](#_Toc165624407)

[Model Results 23](#_Toc165624408)

[Neural Network Model 25](#_Toc165624409)

[Background 25](#_Toc165624410)

[Model Specifications 26](#_Toc165624411)

[Model Building Process 27](#_Toc165624412)

[Model Results 29](#_Toc165624413)

[Conclusion 30](#_Toc165624414)

[References 31](#_Toc165624415)

# Introduction

Communities nationwide face escalating challenges related to crime, substance abuse, and public safety. In recent years, alarming increases in property and violent crimes, alongside persistent issues such as substance abuse and firearm-related fatalities, have underscored the urgent need for proactive interventions.

Our project, "Predict-Align-Prevent," addresses this imperative by harnessing the power of predictive analytics and community-centered interventions to fortify community health and safety initiatives. Collaborating closely with the Public Science Collaborative (PSC), we aim to empower communities with actionable insights derived from robust data analytics.

This report presents a comprehensive analysis of call service data from March 2018 to December 2023 across Mason City, Clear Lake City, and Cerro Gordo County. Through exploratory data analysis and the implementation of predictive models, we seek to uncover trends, identify risk factors, and outline strategies to guide community health and safety initiatives.

# Motivation and Background

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.

# Data Overview

Our analysis is rooted in Cerro Gordo County's comprehensive call for service data, encompassing the period from March 2018 to 2023. This dataset, which underwent rigorous cleaning by our client, provides a robust foundation for understanding community safety dynamics.

Notably, the dataset includes calls from three distinct jurisdictions:

* Mason City Police Department
* Clear Lake Police Department
* Cerro Gordo County Sheriff's Office

Below is a summary of calls across the three distinct jurisdictions.

|  |  |
| --- | --- |
| **Jurisdiction** | **Count** |
| Mason City PD | 109623 |
| Clear Lake PD | 42811 |
| Cerro Gordo County Sherriff | 23187 |

These agencies collectively contribute to a holistic view of public safety concerns within the county.

Key variables within the dataset include:

* Call for Service Type and Category: The dataset encompasses various types of calls, each categorized based on the nature of the reported incident. Understanding these classifications is crucial for identifying patterns and trends in community safety concerns.
* Temporal Elements: Time-related variables, such as timestamps, allow us to discern temporal patterns in call volume and incident occurrence. Analyzing trends over different time intervals provides insights into peak activity periods and seasonal variations in community safety dynamics.
* Geospatial Elements: Geographical attributes, including location data, enable spatial analysis to identify hotspots of activity and areas with heightened risk. By mapping incident locations, we can visualize spatial trends and prioritize resource allocation for targeted interventions.

Below is a summary of types of calls in the dataset.

|  |  |
| --- | --- |
| **Category** | **Count** |
| Other | 140619 |
| Nuisance | 12256 |
| Health | 11943 |
| Domestic | 3740 |
| Blight | 3451 |
| Violence | 2166 |
| Substance-Off Premise | 984 |
| Substance-Driving | 366 |
| Substance-On Premise | 96 |

The categories are defined as such:

On-Premise Substance: This category captures substance-related calls that are directly related to an substance retailer, including liquor and tobacco compliance checks and violations.

Off-Premise Substance: This category includes substance-related calls that are not necessarily directly tied to a specific alcohol or tobacco retailer, such as public intoxication, underage liquor violations, drug activity and complaints, and tobacco law violations.

Substance-Driving: This category includes intoxicated driver calls and OWI checks.

Blight: This category captures calls that may be associated with neighborhood blight, including damaged property, vandalism, abandoned vehicles, illegal dumping, and littering.

Health: This category includes health-related calls, such as those labeled as medical, as well as welfare checks of individuals.

Nuisance: This category contains calls that indicate possible nuisances to neighbors, including disturbances, noise complaints, parties, and disputes.

Violence: This category captures calls relating to physical violence, including assault, robbery, sexual assault, and armed subjects

Domestic: This category includes verbal and physical domestic calls for service.

Through meticulous exploration and analysis of these important variables, we aim to uncover actionable insights that inform our predictive modeling efforts and guide community-centered interventions.

# Exploratory Data Analysis

Exploratory Data Analysis was conducted on the call for service data as previously described using R code.

A graph of different sources

Description automatically generated

Figure 1: Distribution of the call service data among the 3 kinds of distinctions: Mason City PD, Clear Lake PD and Cerro Gordo County Sherriff.

A graph showing a number of colored bars

Description automatically generated

Figure 2: Distribution of the call service data across the months of the year by weekday.

A graph of different colored lines

Description automatically generated

Figure 3: Time series of service calls over hours of the day by years in data set.

A graph showing the number of months

Description automatically generated

Figure 4: Time series of service calls over months of the years in data set.

A graph showing different colored squares

Description automatically generated with medium confidence

Figure 5: Distribution of the types of calls across weekend nighttime, weekend daytime, weekday nighttime, and weekday daytime.

A graph of different colored bars

Description automatically generated

Figure 6: Distribution of type of call made by time of day categorized in weekday daytime, weekday nighttime, weekend daytime, or weekend nighttime.

A graph of different colored lines

Description automatically generated

Figure 7: Time series of Substance-Driving, Substance-Off Premise, Substance-On Premise and Violence category calls by hour of day for the years in the dataset.

A graph of different colored bars

Description automatically generated

Figure 8: Distribution of substance related calls by weekdays and call source.

A graph showing different colored squares

Description automatically generated

Figure 9: Distribution of the Substance-Off Premise calls by month.

A graph of different colored lines

Description automatically generated

Figure 10: Time series of Substance-Off Premise calls for all years in the dataset. Black line represents the trend-line of the data for this category.

A screenshot of a map

Description automatically generated

Figure 11: Heat map of substance off premise service calls at 05:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 12: Heat map of substance off premise service calls at 10:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 13: Heat map of substance off premise service calls at 21:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 14: Heat map of substance driving service calls at 05:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 15: Heat map of substance driving service calls at 10:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 16: Heat map of substance driving service calls at 21:00 hours within Cerro Gordo County

A screenshot of a map

Description automatically generated

Figure 17: Heat map of substance on premise service calls within Cerro Gordo County at 10:00 hours

A screenshot of a map

Description automatically generated

Figure 18: Heat map of substance on premise service calls within Cerro Gordo County at 21:00 hours

A screenshot of a map

Description automatically generated

Figure 19: Heat map of Nuisance service calls within Cerro Gordo County at 08:00 hours

A screenshot of a map

Description automatically generated

Figure 20: Heat map of Nuisance service calls within Cerro Gordo County at 21:00 hours

A screenshot of a map

Description automatically generated

Figure 21: Heat map of Hit and Run or Theft or Shoplifting service calls within Cerro Gordo County at 08:00 hours

A screenshot of a map

Description automatically generated

Figure 22: Heat map of Hit and Run or Theft or Shoplifting service calls within Cerro Gordo County at 21:00 hours

A screenshot of a map

Description automatically generated

Figure 23: Heat map of Violence related service calls within Cerro Gordo County at 08:00 hours

A screenshot of a map

Description automatically generated

Figure 24: Heat map of Violence related service calls within Cerro Gordo County at 21:00 hours

A map of different colored squares

Description automatically generated

Figure 25: Geospatial distribution of calls for service between Clear Lake (left), Mason City (right), and Cerro Gordo (larger blocks). Levels of call frequency is as follows from highest to lowest: yellow, green, blue, purple, white

A graph of different colored lines

Description automatically generated

Figure 26: Time series of Health calls for all years in the dataset. Black line represents the trend-line of the data for this category.

A graph showing the days of the week

Description automatically generated

Figure 27: Distribution of Theft and Shoplifting calls by days of the week for the years in the data.

A graph showing the number of shoplifters

Description automatically generated

Figure 28: Time series of Theft and Shoplifting calls from the Other category for all years in the dataset. Black line represents the trend-line of the data for this category.

# High-Level Summary of EDA

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.

# Logistic Regression Model

## Background

Logistic regression is a statistical technique used for binary classification tasks, where the outcome variable is categorical and has two possible outcomes. It’s widely employed due to its simplicity, interpretability, and effectiveness in modeling relationships between independent variables and the probability of a particular outcome.

Unlike linear regression, which predicts continuous outcomes, logistic regression predicts the probability of a binary outcome. A threshold is then applied to classify observations into one of the two classes based on the predicted probabilities.

## Model Specifications

Understanding the temporal patterns and trends in theft incidents is important for the public to allocate resources effectively, plan proactively, and make informed decisions. To address this, we've developed a Logistic Regression model focused on predicting theft or shoplifting calls.

Data:

We've curated a comprehensive dataset from March 2018 to December 2022, ensuring our model captures recent trends and patterns in theft incidents. This timeframe enables us to generate accurate predictions for future periods.

Features Used:

1. Block Group GEOID: This feature denotes the specific grid where the call for service occurred. It serves as a vital predictor in estimating the frequency of theft calls per hour.
2. Hour of the Day (cfs\_hour): Identifying the hour of the day is crucial for discerning peak hours and periods of heightened activity in theft incidents. This temporal feature aids in understanding the dynamics of theft occurrence throughout the day.
3. Day of the Week (cfs\_day): Including the day of the week allows our model to capture any weekly patterns or variations in theft call volume. This feature enhances the model's ability to detect recurring trends over different days.

The logistic regression model aims to provide accurate predictions of theft calls per hour. These predictions offer valuable insights into temporal dynamics and trends in theft incidents, empowering public safety agencies to make data-driven decisions and allocate resources efficiently.

## Model Building Process

Through a systematic approach, I crafted a robust Logistic Regression model tailored to predict Theft calls per hour. Here's how I customized the process to achieve this:

Data Preparation:

I split the dataset into training and test sets, allocating 70% for training and 30% for testing. This ensured a balanced evaluation of the model's performance.

Logistic Model:

At first, the logistic regression model was built with default parameters, serving as a foundation. This foundational model provided a baseline for optimization. After looking at the summary of the model, the next step was removing necessary predictors to lead to a better model.

Performance Evaluation:

After assessing the model's performance using metrics, including accuracy, f1 Score, and confusion matrix. These metrics highlighted the model's predictive power.

Evaluation:

After examining the final logistic regression model's accuracy to see how well it could predict Theft calls per hour. Accuracy tells us how often the model's predictions are correct, which is important.

By following a step-by-step process to build this model, I aimed to create a strong Logistic Regression model that's good at predicting Theft calls. Being able to predict this helps us figure out where to put resources and make better decisions to keep people safe. It makes our plans more effective and helps us respond better to emergencies.

## Model Results

The logistic regression model aimed to predict theft or shoplifting based on features including Block Group GEOID, Hour of the Day (cfs\_hour), and Day of the Week (cfs\_day). After data preparation, model building, and performance evaluation, the final model achieved an accuracy of 0.68 and an F1 score of 0.81 which is better than the original model which obtained an accuracy of 0.61 and an F1 Score of .75.

This suggests that the model accurately anticipates whether a theft or shoplifting call will happen around 68% of the time. Although accuracy is a valuable measure for assessing model performance, it might not fully capture the effectiveness in datasets where one class (such as no theft or shoplifting call) outweighs the other. A score of 0.81 indicates that the model maintains a favorable equilibrium between precision (its capability to avoid mislabeling negative samples as positive) and recall (its ability to detect all positive samples).

# Random Forest Model

## Background

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, which combines multiple individual models to produce a more accurate and robust final model.

At its core, Random Forest 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.

Literature review provided valuable insights into the application of Random Forest models in analyzing crime patterns based on spatial and temporal attributes. For instance, studies such as “Incorporating Space and Time into Random Forest Models for Analyzing Geospatial Patterns of Drug-Related Crime Incidents in a Major U.S. Metropolitan Area” (Xia, et.al, 2021) have demonstrated the effectiveness of Random Forest in capturing spatiotemporal patterns in crime data, particularly in urban areas like Chicago. These studies incorporated spatial and temporal lag variables to detect dependencies and relationships and utilized model optimization techniques such as guided regularized Random Forest and grid search to enhance predictive capabilities.

Furthermore, “Predicting Crime Using Time and Location Data” (Yuki, et.al, 2019) and “Crime Prediction Using Spatio-Temporal Data” (Hossain, et.al, 2020) showcased the importance of feature selection and model optimization techniques in enhancing the predictive capabilities of Random Forest models when dealing with imbalanced datasets and incorporating both spatial and temporal factors.

In the following sections, we will delve into the specifics of how Random Forest models were trained, optimized, and evaluated to analyze crime patterns based on temporal attributes in our dataset.

## Model Specifications

The primary objective of the Random Forest Regression model is to predict the number of [category] 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.

Data Manipulation

* Filtered Call for Service (CFS) Data: the dataset used for model training and prediction spans from the years March 2018 to December 2022. This timeframe ensures that the model captures recent trends and patterns in [category] calls, enabling accurate predictions for future time periods.
* Narrowed Category: the focus of the analysis is narrowed to category-specific calls, ensuring that the model’s predictions are specific to the target incident type.

Features Used:

1. cfs\_type: the type of call for service provides valuable information about the nature of the incident, which is crucial for predicting the number of [category] calls per hour.
2. cfs\_year: incorporating the year allows the model to capture potential temporal trends and variations in [category] incidents over time.
3. cfs\_month: the month of the year provides seasonal context, enabling the model to account for seasonal fluctuations in [category] call volume.
4. cfs\_hour: the hour of the day is a critical temporal feature, as it helps identify peak hours and periods of heightened activity for [category] incidents.
5. cfs\_day: including the day of the week allows the model to capture any weekly patterns or variations in [category] call volume.

By leveraging these features, the Random Forest Regression model aims to accurately predict the number of category-specific calls per hour, providing valuable insights into temporal dynamics and trends in [category] incidents.

## Model Building Process

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.

## Model Results

To start, the following shows the variable importance in the model that was developed. Cfs\_day and cfs\_hour have the highest significance in predicting the number of calls per hour.

A graph with text on it

Description automatically generated

The analysis focused on the Nuisance category within the police call for service data. Random Forest models were constructed and evaluated to predict the number of calls in each hour.

The original Random Forest model yielded the following performance metrics. Upon hyperparameter tuning, the model’s performance showed slight variations. Furthermore, employing bagging techniques to the tuned Random Forest model resulted in the following.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **Tuned Model** | **Bagged Model** |
| **Mean Absolute Error** | 0.2152 | 0.2210 | 0.2212 |
| **Mean Squared Error** | 0.1275 | 0.1334 | 0.1332 |
| **R2** | 0.0202 | 0.0110 | 0.0115 |
| **Out of Bag Error** |  |  | 0.1109 |

The out-of-bag error rate of 11.09% suggests that, on average, the bagged model misclassifies approximately 11.09% of Nuisance category observations in the training data that were not used in the construction of the trees making the prediction.

Analysis was extended to focus on the Theft or Shoplifting category within the Other category of the police call for service data. Random Forest models were constructed and evaluated to predict the number of calls in each hour.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **Tuned Model** | **Bagged Model** |
| **Mean Absolute Error** | 0.1997 | 0.1997 | 0.2000 |
| **Mean Squared Error** | 0.1230 | 0.1230 | 0.1222 |
| **R2** | 0.0096 | 0.0096 | 0.0136 |
| **Out of Bag Error** |  |  | 0.0978 |

The out-of-bag error rate of 9.78% indicates that, on average, the bagged model misclassifies approximately 9.78% of the Theft or Shoplifting category observations in the training data that were not used in the construction of the trees making the prediction.

Furthermore, analysis shifted focus to the Hit and Run Collisions category within the Other category of the police call for service data. Random Forest models were constructed and evaluated to predict the number of calls in each hour.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **Tuned Model** | **Bagged Model** |
| **Mean Absolute Error** | 0.0721 | 0.0721 | 0.0705 |
| **Mean Squared Error** | 0.0310 | 0.0310 | 0.0307 |
| **R2** | 0.0006 | 0.0006 | 0.0017 |
| **Out of Bag Error** |  |  | 0.0455 |

The out-of-bag error rate of 4.55% suggests that, on average, the bagged model misclassifies approximately 4.55% of Hit and Run Collisions category observations in the training data that were not used in the construction of the trees making the prediction.

These results indicate a model predictive performance of the Random Forest models in forecasting the number of calls related to the Nuisance, Theft or Shoplifting, and Hit and Run Collisions category. Further refinement and exploration of alternative modeling techniques may be necessary to enhance prediction accuracy and reliability.

# Neural Network Model

## Background

A basic Neural network is based on a linear Perceptron model. A perceptron model works by going through an algorithm where it starts off by randomly assigning weights to the features. Then it runs the model and uses a sign function to classify samples into the categories, it assesses the accuracy of the model, and then chooses a random misclassified sample and updates the weights because of it. This happens until convergence is reached or a max iteration is set. For this model, that max iteration was 200 for efficiency purposes.

A Neural network model builds on this, by connecting many perceptron models in different layers, and instead of using a sign function it uses a more complex function. For this project, a ReLu function was used due to the gradient vanishing problems that can come at outlier values of which this dataset has plenty of.

A convolutional neural network (CNN) builds off a basic neural network by incorporating specialized layers for spatial hierarchies in data. While basic neural networks process each input feature independently, CNNs use convolutional layers to detect patterns across neighboring features. This allows CNNs to efficiently recognize spatial patterns like edges and textures. Additionally, CNNs often include pooling layers to downsample feature maps and reduce computational complexity. Overall, CNNs are designed to effectively process and extract features from structured data like images, making them well-suited for more complex tasks that may include lags or even spatial data.

A recurrent neural network (RNN) builds upon a basic neural network by incorporating feedback loops within the network, allowing it to capture sequential dependencies in the data. Unlike basic neural networks, RNNs have connections that form directed cycles, enabling them to process sequences of inputs. RNNs are particularly well-suited for forecasting tasks because they can retain memory of past inputs, making them adept at modeling time series data. This ability to capture temporal dependencies gives RNNs an advantage over convolutional neural networks (CNNs) for forecasting tasks where the order of input data is crucial, such as predicting future values in a time series.

There has been some previous work in this field that was referenced. For example, in Pittsburgh, there was some mapping done using an Artificial Neural Networks. Some similarities between that project and ours were that we both used feed-forward neural networks to predict crimes using time series data. A large difference between that model and this project was that for that project they focused more on spatial data, whereas in this project, the focus was more on temporal data (Olligschlaeger).

Then we referenced an article that did crime hotspot forecasting with spatial and temporal information. This used a recurrent neural network, but it did forecasting which aligned with the goals of this project (Zhuang).

In the following sections we will go into how we implemented a neural network to the data of Cerro Gordo County, Iowa.

## Model Specifications

The goal of the neural network model was to predict the number of service calls for a given hour of a given day. This will in turn help the police department in staffing, response times, and availability during those times in the hotspot regions.

Data Manipulation

* Filtered Call for Service (CFS) Data: the dataset used for model training and prediction spans from the years March 2018 to December 2022. This timeframe ensures that the model captures recent trends and patterns in [category] calls, enabling accurate predictions for future time periods.
* Then further manipulation was done to set the data up for neural network analysis including scaling the data to input into the model frameworks. Then, we also experimented with subsetting the data to get more tailored insights for the user based on the type of service call.

Features Used:

1. cfs\_type: the type of call for service provides valuable information about the nature of the incident, which is crucial for predicting the number of [category] calls per hour. Subsetted data based on this
2. cfs\_year: incorporating the year allows the model to capture potential temporal trends and variations in [category] incidents over time.
3. cfs\_month: the month of the year provides seasonal context, enabling the model to account for seasonal fluctuations in [category] call volume.
4. cfs\_hour: the hour of the day is a critical temporal feature, as it helps identify peak hours and periods of heightened activity for [category] incidents.
5. cfs\_day: including the day of the week allows the model to capture any weekly patterns or variations in [category] call volume.
6. Lags: Incorporating lag variables allows for a forecasting model by looking at past data as it relates to the data of the given row

## Model Building Process

1. Preprocess the Data:
   * Handle missing values: Impute missing data or remove instances with missing values.
   * Normalize or standardize numerical features to ensure all features have the same scale.
   * Encode categorical variables into numerical representations (e.g., one-hot encoding).
   * Split the data into training (.8), validation (.2), and testing sets (done with training data).
   * Encode categorical variables into numerical representations (e.g., one-hot encoding).
2. Make Single Layer Linear Perceptron Model
   * Use Perceptron Learning Algorithm to predict whether there will be a service call on a given day
   * Produce results that will be used in overall Neural
3. Design the Neural Network Architecture:
   * Determine the type of neural network suitable for the problem, started with MLP
     + MLP Extension: Multilayer Perceptron (MLP) extends perceptron by introducing hidden layers, facilitating learning of complex, non-linear relationships through weighted connections.
     + Numerical Prediction: MLPs are adept at predicting numerical outcomes, employing output neurons to directly produce predicted values, suitable for regression tasks such as predicting number of calls for each hour.
     + Backpropagation Learning: MLPs adjust weights and biases via backpropagation, iteratively optimizing predictions by minimizing the difference between predicted and actual values during training.
   * Convolutional Networks Extension: Convolutional Neural Networks (CNNs) build upon MLPs by incorporating convolutional layers, enabling automatic feature extraction through local receptive fields.
   * Recurrent Neural Networks
     + Temporal Dependency Handling: Recurrent Neural Networks (RNNs) extend from MLPs and CNNs by incorporating recurrent connections, enabling them to capture temporal dependencies in sequential data such as time series.
     + Dynamic Memory Utilization: RNNs utilize internal memory units to store information about past observations, allowing them to effectively model sequential patterns and make accurate predictions in time series data.
   * Decide the number of layers and neurons in each layer.
   * Choose activation functions for hidden layers (e.g., ReLU, sigmoid, tanh).
   * Select an appropriate loss function based on the problem (e.g., mean squared error for regression, cross-entropy for classification).
   * Define the optimizer (e.g., SGD, Adam) and its parameters.
   * Set any other hyperparameters such as learning rate, batch size, and regularization techniques.
4. Train the Neural Network:

* Feed the training data into the model and adjust the weights iteratively to minimize the loss function.
* Monitor the model's performance on the validation set to prevent overfitting.
* Experiment with different hyperparameters if necessary to improve performance.
* Stop training when the model performance on the validation set starts to degrade or after a fixed number of epochs.

1. Evaluate the Model:
   * Assess the model's performance on the test set using appropriate metrics (e.g., MSE, MAE, F1-score, RMSE).
   * Analyze any discrepancies between the model's predictions and the actual values.
   * Consider additional evaluation techniques such as cross-validation for more reliable performance estimation.

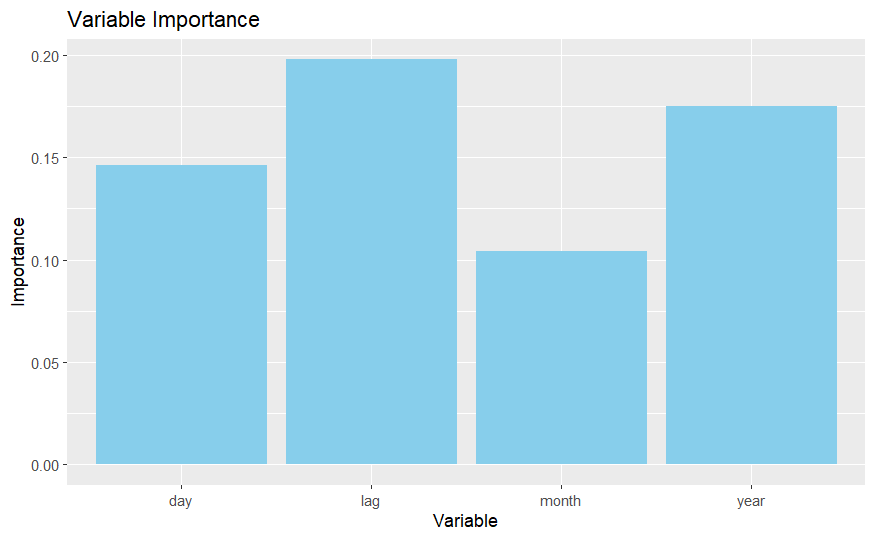
This model building process trains a neural network model that will use multiple layers to predict the number of service calls for a certain category for a given hour of a given day.

## Model Results

For the final model, we focused on violence due to demands from the clients and to achieve more accurate predictions as these models were tried on various subsets of the data.

First, we will show the variable importances of the final neural network model which was made using a recurrent neural network framework.

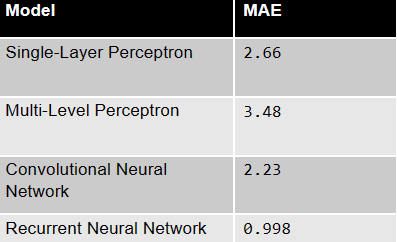
Variable Importance plot:



From this we can see that lag variables are more important in RNN than other types of models. This is because RNNs are designed to capture temporal dependencies in sequential data. Lag variables directly capture the temporal relationship between past observations and current predictions, making them highly relevant in modeling time series data. RNNs also have an internal memory mechanism that allows them to retain information from past time steps. Lag variables provide direct access to historical data, which is crucial for making accurate predictions in time series forecasting tasks.

To show that RNN, had the best metrics, we can look at this table.

Model summary table:



A RNN can give the best MAE for time series analysis, but future work can include further tuning of all of the models.

# Conclusion

Our analysis of police call for service data has yielded several noteworthy findings, shedding light on the temporal patterns and influential factors underlying incident reporting. Across the years 2018-2023, we observed a consistent peak frequency of calls during the months of May to August, indicative of heightened activity during the summer months. Additionally, our analysis revealed a notable spike in call frequency around 4 PM throughout the days, suggesting a peak period of incident reporting during late afternoon hours.

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\_group\_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\_day and cfs\_hour as the most influential predictors, 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.

Moving forward, our next steps involve expanding the scope of our modeling efforts. We plan to develop new models utilizing Logistic Regression, Random Forest, and Neural Networks. These new models will aim to predict different features within the dataset, offering a more comprehensive understanding of the underlying patterns and dynamics of police call for service data.

Furthermore, the need for additional data collection is paramount. Certain categories within our dataset contained limited instances, potentially impacting the robustness of our models. Therefore, collecting more data, especially for these underrepresented categories, will be essential for improving the accuracy and reliability of our predictive models.

Moreover, an intriguing avenue for future investigation involves exploring the impact of police shifts on incident reporting. We intend to analyze the correlation between police shifts and incident reporting to discern how staffing levels influence reported incidents. Understanding these dynamics will provide valuable insights into the factors driving fluctuations in incident counts, shedding light on the underlying mechanisms behind the observed variability in the data.

In essence, our future endeavors will continue to advance our understanding and predictive capabilities in leveraging police call for service data to enhance safety and decision-making processes.

# References

Hossain, S., Abtahee, A., Kashem, I., Hoque, M.M., Sarker, I.H. (2020). Crime Prediction Using Spatio-Temporal Data. In: Chaubey, N., Parikh, S., Amin, K. (eds) Computing Science, Communication and Security. COMS2 2020. Communications in Computer and Information Science, vol 1235. Springer, Singapore. https://doi.org/10.1007/978-981-15-6648-6\_22

Jesia Quader Yuki, Md. Mahfil Quader Sakib, Zaisha Zamal, Khan Mohammad Habibullah, and Amit Kumar Das. 2019. Predicting Crime Using Time and Location Data. In Proceedings of the 7th International Conference on Computer and Communications Management (ICCCM '19). Association for Computing Machinery, New York, NY, USA, 124–128. <https://doi.org/10.1145/3348445.3348483>

Olligschlaeger, A. M. (1997). Artificial neural networks and crime mapping. Crime mapping and crime prevention, 1, 313.

Xia Z, Stewart K, Fan J. Incorporating space and time into random forest models for analyzing geospatial patterns of drug-related crime incidents in a major U.S. metropolitan area. Comput Environ Urban Syst. 2021 May;87:101599. doi: 10.1016/j.compenvurbsys.2021.101599. Epub 2021 Jan 29. PMID: 33828350; PMCID: PMC8021089.

Zhuang, Y., Almeida, M., Morabito, M., & Ding, W. (2017, August). Crime hot spot forecasting: A recurrent model with spatial and temporal information. In 2017 IEEE International Conference on Big Knowledge (ICBK) (pp. 143-150). IEEE.