Predicting Family Violence from APD Crime Reports and Location

Callihan Bertley, Bhanu Sharma, Kashif Ghani, Anthony Truong, Brian Huynh, Adam Wilensky

Introduction

Content Warning: This project involves discussions of sexual assault/violence and domestic violence. If you or someone you know is experiencing these issues, please seek help. National Sexual Assault Hotline: 1-800-656-4673; National Domestic Violence Hotline: 1-800-799-7233.

Objective

This report delves into an analysis of crime data from Austin, Texas, for the year 2022, with a specific emphasis on family violence and related criminal activities. It seeks to provide insights into the patterns and dynamics of these crimes in the rapidly evolving urban landscape of Austin.

Background and Motivation

Austin has recently experienced a significant influx of new residents, leading to increased urban congestion and diverse demographic changes (2). These dynamics can have complex effects on the nature and frequency of crimes, particularly those related to family violence. This project is motivated by the need to understand how these evolving urban conditions influence criminal activities and to identify specific patterns that emerge within this unique context.

Importance of the Problem

The study of crime in Austin, specifically in the context of its changing demographics and increasing urban density, is crucial for several reasons:

1. **Adapting to Demographic Changes:** The recent surge in immigration and population growth in Austin necessitates a fresh look at crime data to understand how these changes impact the nature and frequency of family violence and other crimes.
2. **Informing Local Policies and Initiatives:** Insights from this analysis can guide policymakers, law enforcement, and community leaders in Austin to develop tailored strategies that address the specific challenges of a growing and diversifying city.
3. **Community Awareness and Support:** By shedding light on the current state of family violence in Austin, this report aims to foster community awareness and encourage the development of targeted support systems for affected individuals.
4. **Broader Societal Implications:** While focused on Austin, this study's findings can provide valuable lessons for other cities experiencing similar demographic and urban shifts.

This report will explore the 2022 crime data for Austin, Texas, with a focus on family violence, employing a range of data science techniques to uncover key trends and insights relevant to the city's crime-scape centered on family violence.

Data

Objective

The objective of this section is to describe the dataset utilized for analyzing crime and family violence in Austin, Texas, specifically focusing on the data from 2022 up to the present year 2023. This part also details the data preparation and cleaning processes.

Source of the Data

The dataset was sourced from the public crime records of Austin, Texas (1). It contains detailed information about various criminal incidents reported within the city, providing a comprehensive view of the crime landscape.

Features/Variables in the Dataset

Initially, the dataset included 27 columns, such as:

- **Incident Number**: Unique identifier for each crime incident.
- **Highest Offense Description and Code**: Details of the primary offense in the incident.
- **Family Violence**: Indicator of family violence involvement.
- **Occurred Date and Time**: Timestamps of the crime occurrence.
- **Geographical Details**: Including Location Type, Address, Zip Code.
- **Jurisdictional Information**: Such as Council District, APD Sector, APD District.
- **Investigative Details**: Including Census Tract, Clearance Status, Clearance Date.
- **Crime Categorization**: UCR Category, Category Description.
- **Geospatial Data**: Coordinates including X, Y, Latitude, Longitude.

Data Cleaning and Preprocessing

1. **Removing Unnecessary Columns**: Non-essential columns such as 'Clearance Date', 'UCR Category', and various location-specific details like 'X-coordinate', 'Y-coordinate' were removed for a more streamlined analysis.
2. **Handling Missing Values**: Rows containing NaN values were eliminated to ensure data integrity.
3. **Data Filtering**: The focus was narrowed to data from 2022 to the present year 2023, providing a current and relevant dataset for analysis.
4. **Data Encoding and Normalization**:
    - Binary encoding was applied to 'Family Violence'.

- Multi-label one-hot and general encoding transformed categorical data into machine-readable formats for variables like 'Highest Offense Code', 'Location Type', and others.
- Date and time columns were processed, including extracting month and day, and normalizing 'Occurred Time'.
- Min-max scaling was used for 'Latitude' and 'Longitude' to maintain consistency.

Exploratory Analysis

Objective

The objective of this section is to delve into the dataset to uncover initial insights about family violence and crime patterns in Austin, Texas. This exploration is guided by visualizations and the formulation of hypotheses, primarily utilizing the Decision Tree Classifier for its high precision, recall, and interpretability.

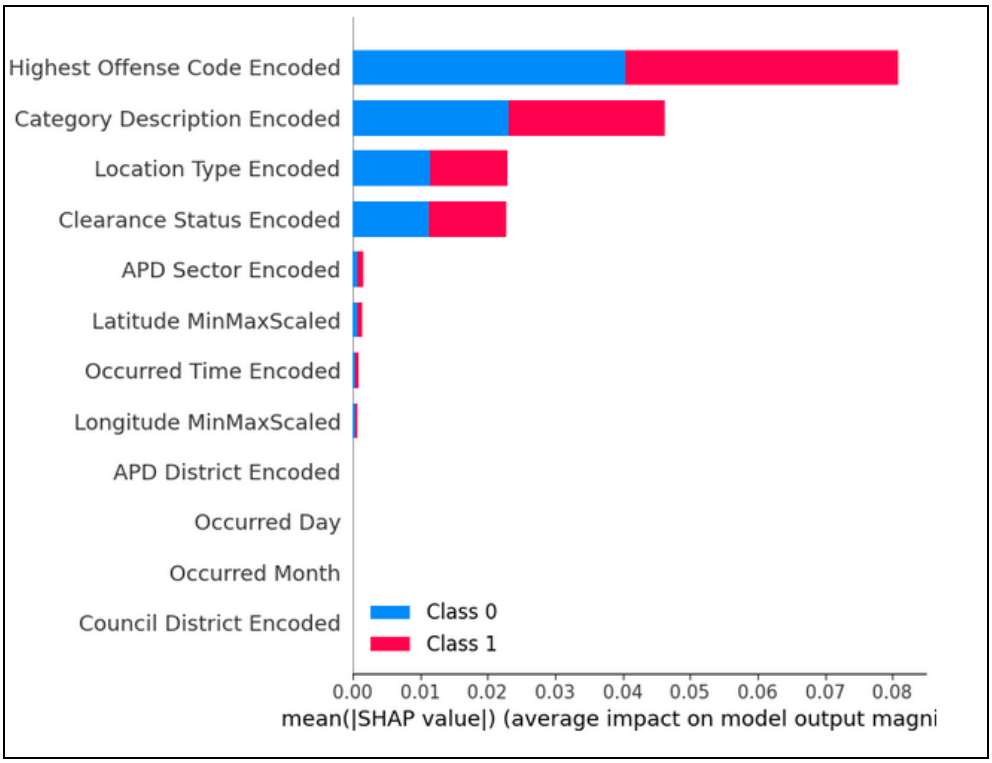
Hypotheses Development

Four hypotheses were developed based on the data exploration and feature importance determined through SHAP (SHapley Additive exPlanations) values from the Decision Tree Classifier:

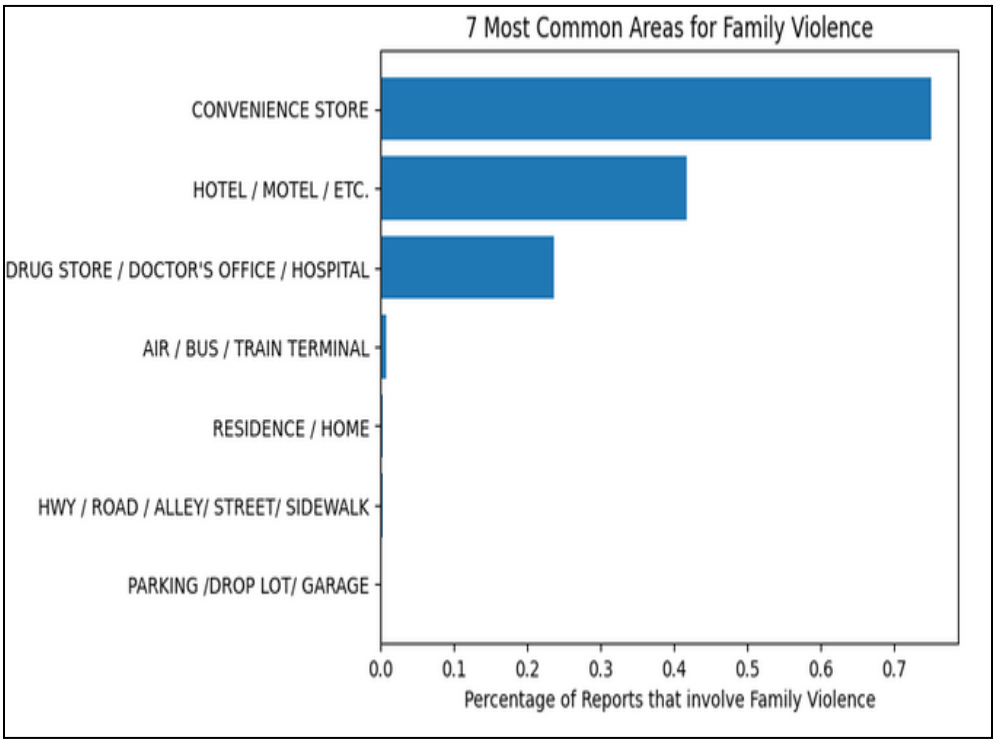
1. **Hypothesis 1: Family Violence Is More Common in Winter**
   Analysis of SHAP values showed that the 'Occurred Month' feature had minimal impact on predicting family violence. This suggests that the season, including winter, does not significantly influence the likelihood of family violence in reported incidents.
2. **Hypothesis 2: Family Violence Occurs More in Areas Away from the City**
   The minimal impact of 'Latitude' and 'Longitude' features on the prediction model implies that the location of an incident in relation to Austin's city center does not significantly affect the probability of the incident involving family violence.
3. **Hypothesis 3: Family Violence Is More Common in Residential Areas**
   Examination of the 'Location Type' feature indicated that only 0.002966% of incidents at 'Residence/Home' involved family violence, which is lower compared to other locations like 'Convenience Stores' and 'Hotels/Motels'. This suggests that family violence is not more prevalent in residential areas.
4. **Hypothesis 4: Incidents Involving Rape Are Less Likely to Include Family Violence**
   Analysis revealed that reports of rape without family violence are more common than those with family violence. This suggests that incidents involving rape are less likely to coincide with family violence.

Visualizations and Analysis

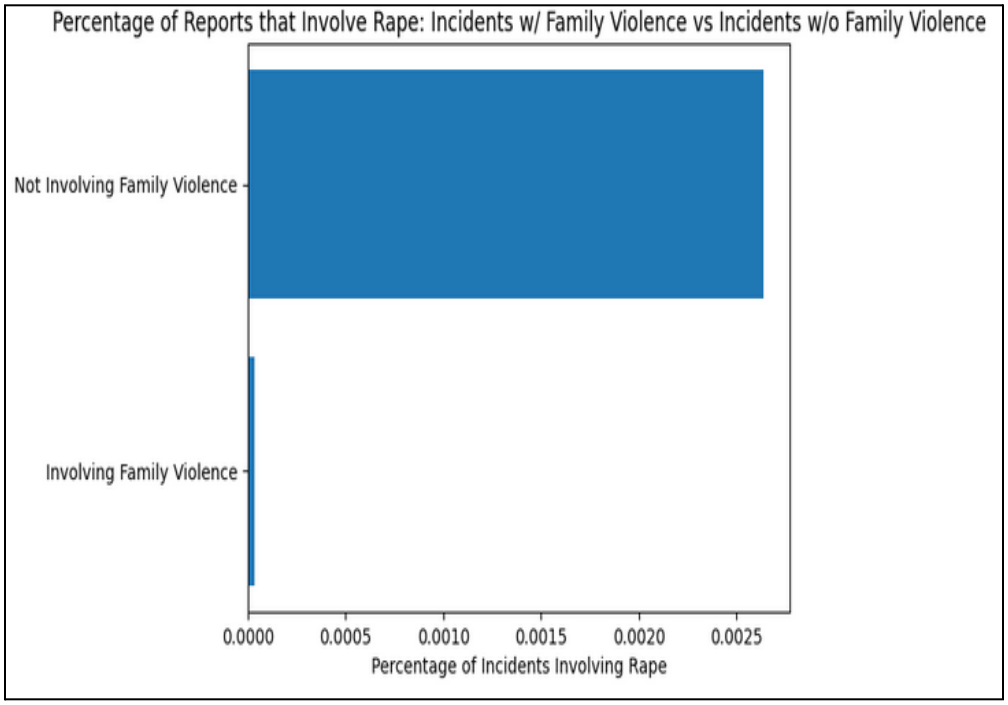
1. **SHAP Summary Plot:** Used for Hypotheses 1 and 2, this plot indicated the average impact of each feature on the model's predictions, guiding the conclusions drawn for these hypotheses.



2. **Horizontal Bar Plot for Family Violence in Different Locations:** This visualization showed the proportion of family violence cases in various location types. The low percentage of such cases in residential areas led to the conclusion for Hypothesis 3.



3. **Bar Plot Comparing Rape Incidents with and without Family Violence:** This plot visually represented the proportion of rape incidents involving family violence versus those that did not, leading to the validation of Hypothesis 4.



Modeling

Objective

The aim of this phase was to apply machine learning techniques for classification purposes in predicting family violence within the crime dataset of Austin, Texas. Various models were evaluated, not just for accuracy, which was consistently high across models, but also for other factors crucial to the study's objectives.

Definition of the Task

The task was approached as a binary classification problem. The primary goal was to predict whether an incident reported in the crime dataset was related to family violence. This classification task falls under supervised learning, utilizing a dataset with clearly defined labels.

Algorithms and Methods Used

Several machine learning models were employed in this phase:

1. Decision Tree Classifier: Selected for its balance between simplicity and predictive power.

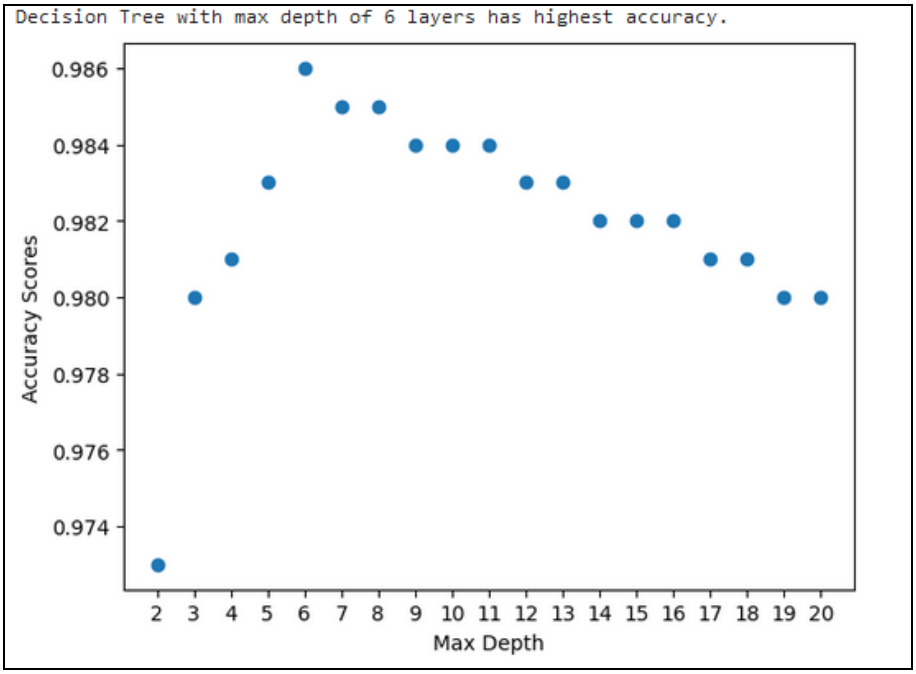
2. Random Forest Classifier: An ensemble approach to gauge performance improvements over a single Decision Tree.
3. Logistic Regression: A fundamental model for binary classification tasks.
4. Naive Bayes: Known for its efficiency in classification tasks.
5. Support Vector Machine (SVM): Tested for its effectiveness in higher-dimensional spaces.
6. K-Nearest Neighbors (KNN): To assess the performance of instance-based learning.
7. Neural Network: Explored for its capability to model complex, non-linear relationships in data.

Problem Setup

The dataset was split into training and test sets in an 80-20 ratio, maintaining a balance between training and evaluation. A random state was set for reproducibility. Features (X) were all the preprocessed columns, excluding the target variable, 'Family Violence', which formed the binary target variable (y).

Selection of the Decision Tree Model

Given that all models showed similar accuracy levels (around 99%), the choice of the Decision Tree Classifier was based on additional criteria beyond accuracy. Crucially, its interpretability and the ease of understanding its decision-making process were significant factors. The optimal depth of the Decision Tree was determined through experimentation, where a depth of 6 layers provided the best balance between model complexity and performance. This depth choice helped in avoiding overfitting while maintaining a high level of prediction accuracy.



Discussion

Objective

The purpose of this section is to discuss the performance of the chosen Decision Tree model, interpret its findings in the context of family violence prediction in Austin's crime data, and acknowledge any limitations or challenges encountered during the analysis.

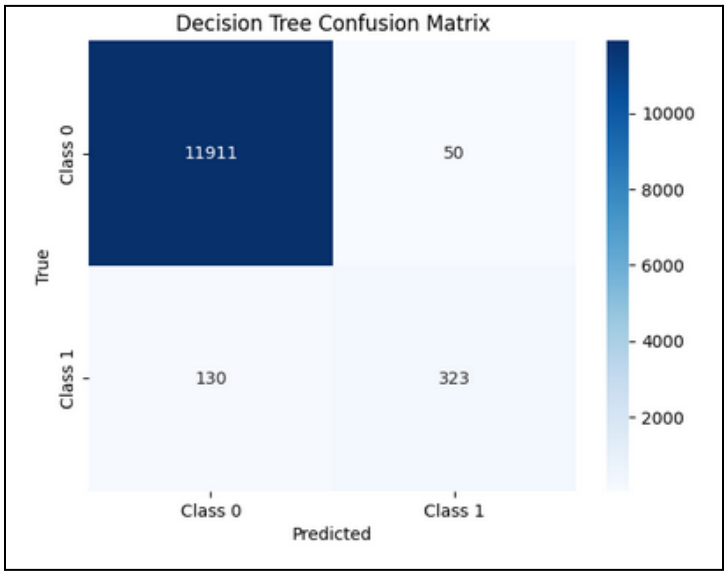
Evaluation Metrics and Results

The Decision Tree Classifier, with a depth of six layers, demonstrated robust performance metrics:

- Accuracy: The model achieved a high accuracy of 99%, indicating its effectiveness in correctly classifying the majority of the cases.
- Precision and Recall for Class 0 (No Family Violence): The model showed a precision of 99% and a recall of 100%, indicating excellent performance in identifying true negatives (cases without family violence).
- Precision and Recall for Class 1 (Family Violence): The precision was 87%, and the recall was 71%. While these values are lower than those for Class 0, they still represent a significant ability of the model to identify true positives (cases with family violence).

Interpretation of the Model's Findings

The model's high accuracy and strong performance in both precision and recall, particularly for Class 0, suggest that it can reliably identify cases without family violence. However, the slightly lower precision and recall for Class 1 indicate some challenges in correctly classifying all true positive cases. This could be attributed to the complexity and varied nature of family violence incidents, which may not always be clearly defined or discernible based on the available features in the dataset.



Potential Limitations and Challenges

1. Imbalanced Dataset: The number of non-family violence cases significantly outnumbered the family violence cases. This imbalance could have influenced the model's ability to learn and predict family violence incidents effectively.
2. Limited Scope of Features: The model's predictions were based on the available features in the dataset. Certain nuances and contextual factors that play a crucial role in family violence incidents might not have been captured.
3. Model Interpretability vs. Complexity: While the Decision Tree's interpretability was a key reason for its selection, this might have come at the cost of capturing more complex patterns that a more sophisticated model might have identified.
4. Generalization Concerns: The model was trained and tested on data specifically from Austin, Texas. Its performance in other regions or contexts might vary, raising questions about its generalizability.

The Decision Tree Classifier, with its high accuracy and reasonable precision and recall, proved to be a useful tool in predicting family violence in the context of Austin's crime data. However, the challenges posed by an imbalanced dataset, the limitations of the available features, and concerns about the model's complexity and generalization are significant limitations. The high number of true negatives indicates that The model is very effective at identifying cases that do not involve family violence. However, the presence of false negatives indicates that the model is more likely to miss actual cases of family violence than to incorrectly identify non-violence cases as such. This could be a critical factor when considering the application of the model, as missing actual cases of family violence (false negatives) could be more detrimental than falsely identifying them (false positives)

Ethics

Objective

Upon reflecting on the completed project on family violence and crime in Austin, Texas, the AREA Plus (4p) Framework prompts a thoughtful retrospective evaluation of the ethical considerations that were integrated into our research. In particular, the principles of **Project Risks**, **Potential Conflicts**, **Stakeholder Involvement**, and **Openness** have significant implications for understanding the ethical dimensions of our work.

With **Project Risks** in mind, it's clear that identifying and mitigating risks at the outset was crucial. In retrospect, examining how we navigated data privacy issues and potential biases in our methodologies can inform the evaluation of our project's success and guide improvements for future research. Understanding how these risks were communicated and managed is crucial for maintaining the integrity of our research and its acceptance by the public and academic community.

Reflecting on **Potential Conflicts** allows us to acknowledge the diverse range of interests and expectations that stakeholders may have held. Assessing how these potential conflicts were handled offers

insight into our project's objectivity and the effectiveness of our engagement strategies. It brings to light the importance of balancing various perspectives while maintaining a clear focus on the research goals.

Engagement with stakeholders is another area where looking back can be particularly instructive. The project's impact and relevance heavily depend on the involvement of those it aims to serve. By evaluating our engagement practices, we can learn valuable lessons about the inclusivity of our approach and the extent to which we successfully incorporated a wide array of voices and experiences into our research process.

Lastly, the principle of **Openness** challenges us to consider how transparent we were in sharing our findings, methodologies, and data. Post-project analysis under this principle could reveal how our commitment to openness may have facilitated the dissemination of knowledge, contributed to public discourse, and encouraged the adoption of best practices within the field.

Conclusion

Our project examining family violence within Austin's crime reports from 2022 onwards yielded a clear finding: certain locations and times have more family violence incidents than others, challenging the assumption that such violence occurs more frequently in residential areas or during specific seasons. The Decision Tree Classifier proved to be an effective, albeit not perfect, predictor of family violence.

For future research, a closer look at socio-economic variables, deeper analysis with real-time data, and a broader scope encompassing similar urban centers could refine our understanding of family violence patterns. The insights gained could guide more targeted and timely interventions, ultimately contributing to safer communities.

The project demonstrates the potential of data-driven approaches to inform public safety and policy, especially in rapidly growing urban settings like Austin. Further research in this vein not only promises to improve local strategies but also offers a template that could be adapted for wider application in comparable cities.

Website

Website for our final project can be found here: http://127.0.0.1:5500/website.html

Acknowledgement

Task	Data Preparation	Graphs	ML	Presentation Prep	Presentation	Report Writing	Website		Total	Contribution
Task %	15	15	15	10	20	15	10		100	
kashif	1	2	3	1	3.5	1	5		16.5	97.05882353
Bhanu	1	2	2	2	8.5	1	0		16.5	97.05882353
Anthony	1	2	1	5	1	6.5	0		16.5	97.05882353
Brian	2	2	3	2	2	0.5	5		16.5	97.05882353
Adam	5	2	2	0	2	6	0		17	100
Callihan	5	5	4	0	3	0	0		17	100
Person G	0	0	0	0	0	0	0		0	0
Total	15	15	15	10	20	15	10	Total	100	
								Max	17	

Bibliography

1. https://data.austintexas.gov/Public-Safety/Crime-Reports/fdj4-gpfu
2. https://www.austinchamber.com/blog/05-09-2023-migration