CMPD Crime Statistics

Cole Bennett
The William States Lee College of Engineering
University of North Carolina at Charlotte
Charlotte, United States of America
nbenne18@charlotte.edu
https://github.com/oopCole/IntroToMachineLearning

I. Introduction

Crime is a major factor in the value of an area. It affects how safe or dangerous a place could be, which determines if a person should live there, or even visit. Predicting crime in an area is important because it helps communities and law enforcement better understand where and when criminal activity is most likely to occur, allowing them to take proactive steps to prevent it. By using past crime data, patrol and surveillance resources can be deployed more efficiently to areas of greatest need. Crime prediction can also improve public safety by reducing response times and helping to protect vulnerable populations. In addition, city planners and policymakers can use these predictions to guide decisions about lighting, transportation, housing, and social services, which can address some of the root causes of crime. When used ethically, crime prediction supports safer neighborhoods and a stronger sense of security for the public.

II. Approach

A. Classification Revisions

Throughout the duration of the project, the goal and scope of the classification task underwent multiple revisions. Several strategies were tested to determine the most effective way to classify crime data while maintaining accuracy and real-world relevance.

a. Binary Classification

The initial approach classified crimes into two categories: violent and non-violent. The Violent-Crime column was used as the ground-truth label, where crimes against a person were classified as violent, and crimes against property or society were classified as non-violent. While this model achieved high accuracy, the results were not particularly useful for practical crime analysis, since the classification was too broad to provide meaningful insight.

b. Crime Classification

The next method attempted to classify each of the 45 individual crime types separately. However, the wide variation in the number of reported cases per crime category caused the model to favor frequently occurring crimes while performing poorly on under-represented ones. This severe class imbalance once again led to reduced accuracy and unreliable predictions.

c. Reduced Class Grouping

To mitigate the effects of data imbalance and reduce model overfitting, the large number of original crime categories was consolidated into a smaller set of broader groups. This reorganization combined related offenses into unified classes, creating a more even distribution of data across categories. As a result, the model was less biased toward high-frequency crimes and demonstrated improved prediction performance for offense types that had previously been underrepresented.

B. Classical Models

Several classical algorithms were tested to determine which produced the most accurate results. Support Vector Classification was initially evaluated but demonstrated very low accuracy and was therefore excluded from further analysis. Naive Bayes and Logistic Regression models were then implemented and compared. Among these, the Naive Bayesian classifier consistently yielded the highest accuracy and was selected as the best-performing classical model. Both classical machine-learning techniques and deep learning models were implemented to evaluate their effectiveness in crime classification. This comparative approach allowed for a direct assessment of the strengths and limitations of traditional statistical methods versus neural-network-based architecture.

C. Deep Learning Models

A spatiotemporal neural network was implemented to model the relationships between crime patterns across both space and time. This architecture was selected due to its ability to handle irregularly occurring events and capture complex patterns in sequential data.

a. 1D Convolutional Layer

A one-dimensional convolutional layer was used to extract spatial features from latitude and longitude data. A kernel size of 3 was selected to detect localized geographic patterns, and the layer generated 32 output feature maps.

b. ReLU Activation Function

A ReLU activation function was applied to introduce non-linearity into the network. This allowed the model to capture more complex patterns within the data while maintaining fast and efficient training performance..

c. LSTM Layer

Incorporated to capture time-based patterns and long-term dependencies within the crime data. This layer helped mitigate the vanishing-gradient problem commonly associated with recurrent networks and produced 50 output features, enabling the model to learn temporal fluctuations in crime activity.

III. Datasets and Training Setup

A. Datasets

a. CMPD Incident Reports and Usage Guidelines

This dataset served as the primary source of crime information for the project. It contains detailed records of reported crime incidents and usage policies that were used to summarize crime statistics and support the classification of offense occurrences.

b. Vulnerability to Displacement by Neighborhood Profile Area (NPA)

This dataset was used to incorporate socioeconomic and community risk factors into the crime analysis. It provided displacement vulnerability indicators at the neighborhood level, allowing the model to account for environmental and demographic conditions related to crime activity.

c. North Carolina Zip Codes by Population (2025)

This dataset was used to quantify population density across different geographic regions. By incorporating population data at the zip code level, crime frequencies could be normalized and better compared across areas with varying population sizes.

B. Training Setup

a. Removal of Missing Data

All records containing missing or incomplete values were removed prior to training to ensure that only valid and reliable data were used in the model. This prevented the learning algorithm from being influenced by corrupted or partial information.

b. Date Field Selection

The fields representing when an incident was reported and when it ended were removed from the dataset. Only the incident start date was retained so that all temporal modeling was based on a consistent time reference.

c. Location Accuracy Filtering

Any records in which the listed address corresponded to the reporting location rather than the true incident location were excluded. This step ensured that the geographic features used by the model accurately reflected where crimes actually occurred.

d. Removal of Unconfirmed Records

All incidents marked as unconfirmed were removed from the dataset to avoid introducing uncertainty or unreliable labels into the training process.

e. Categorical Data Encoding

After filtering, all remaining categorical text fields were converted into integer-based enumerations. This transformation allowed non-numerical features to be properly processed by both classical and deep-learning models.

IV. Results and Analysis

After the text edit has been completed, the paper is ready for th

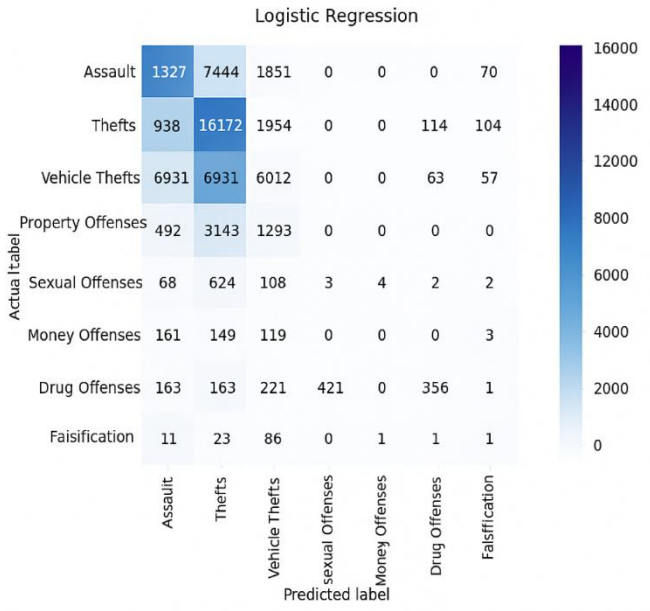

Fig. 1. Logistic Regression Confusion Matrix

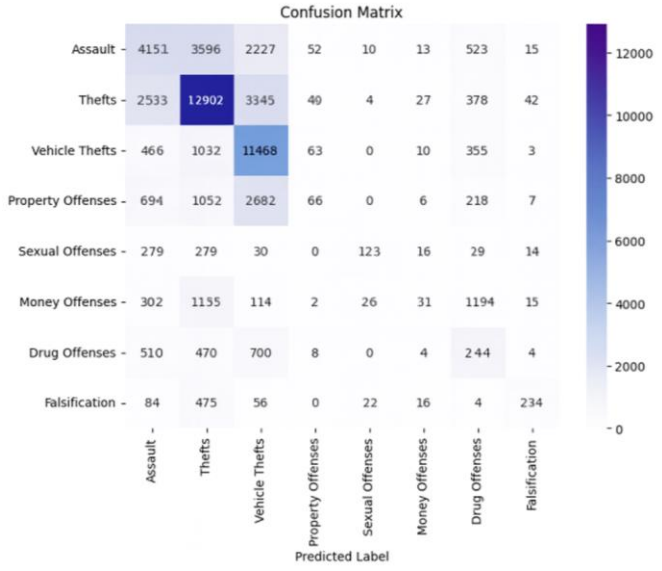

Fig. 2. Spatiotemporal Confusion Matrix

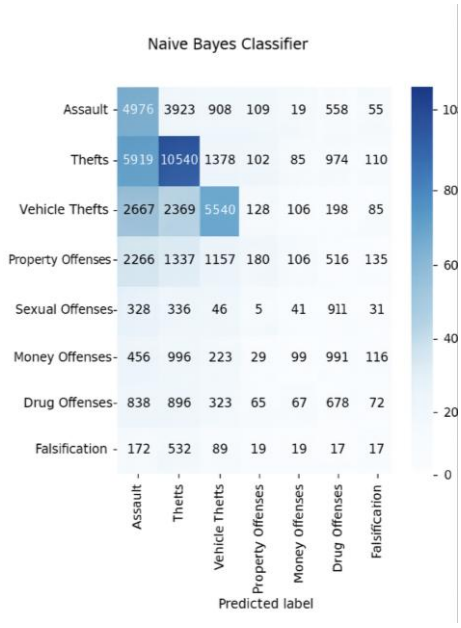

Fig. 3. Naive Bayesian Confusion Matrix

	precision	recall	f1-score	support
0	0.29	0.46	0.35	10682
1	0.48	0.55	0.51	19224
2	0.60	0.44	0.50	13631
3	0.16	0.04	0.06	4912
4	0.51	0.05	0.09	811
5	0.06	0.02	0.03	1720
6	0.17	0.24	0.20	2822
7	0.06	0.02	0.03	935
accuracy			0.41	54737
macro avg	0.29	0.23	0.22	54737
weighted avg	0.41	0.41	0.39	54737

Fig. 4. Naive Bayesian Results

	precision	recall	f1-score	support
0	0.37	0.13	0.19	10682
1	0.42	0.84	0.56	19224
2	0.50	0.44	0.47	13631
3	0.00	0.00	0.00	4912
4	1.00	0.00	0.01	811
5	0.00	0.00	0.00	1720
6	0.07	0.01	0.02	2822
7	0.33	0.00	0.00	935
accuracy			0.43	54737
macro avg	0.34	0.18	0.16	54737
weighted avg	0.37	0.43	0.35	54737

Fig. 5. Logistic Regression Results

	precision	recall	f1-score	support
0	0.46	0.39	0.42	10590
1	0.61	0.66	0.64	19409
2	0.55	0.84	0.66	13605
3	0.28	0.01	0.03	4924
4	0.70	0.17	0.27	749
5	0.22	0.01	0.02	1700
6	0.44	0.41	0.42	2868
7	0.62	0.26	0.37	892
accuracy			0.55	54737
macro avg	0.48	0.35	0.35	54737
weighted avg	0.52	0.55	0.51	54737

Fig. 6. Spatiotemporal Results

These results highlight the importance of evaluating multiple models in order to make meaningful comparisons between prediction accuracy and overall performance. The deep learning models demonstrated superior accuracy and will therefore be emphasized in the final report. The findings also indicate that crime categories with a higher number of reported incidents are predicted more accurately than those with fewer occurrences. Additionally, the features included in the dataset appear to have a stronger influence on predicting crimes such as theft and assault, while offering less predictive power for offenses such as money related and drug-related crimes.

V. Lessons Learned

The main takeaway from the project was that the deep learning approach produced significantly better accuracy and overall performance compared to classical machine-learning models. As the project progressed, the classification goals and methods evolved several times, highlighting the need for adaptability and thoughtful problem solving in machine learning workflows.

Some crime categories were predicted far more accurately than others, emphasizing how essential a balanced dataset is. The most time-intensive portion of the work involved constructing a reliable dataset and performing thorough preprocessing, reinforcing the idea that high-quality data is crucial for building a strong model. One feature that notably influenced the results was race, which had been simplified into a white vs. non-white category. This could prove to be unethical engineering if used slightly incorrectly. I could also distort an entire model.

References

[1] [1] City of Charlotte, “CMPD incidents,” Charlotte Open Data Portal. [Online]. Available: https://data.charlottenc.gov/datasets/charlotte::cmpd-incidents-1/about

[2] [2] City of Charlotte, “Vulnerability to displacement by neighborhood profile area (NPA),” Charlotte Open Data Portal. [Online]. Available: https://data.charlottenc.gov/datasets/charlotte::vulnerability-to-displacement-by-npa/about

[3] [3] North Carolina Demographics, “North Carolina zip codes by population,” 2025. [Online]. Available: https://www.northcarolina-demographics.com/zip_codes_by_population