Predicting Crime in Charlotte

Aidan Cowan
The William States Lee College of Engineering
University of North Carolina at Charlotte
Charlotte, United States of America
acowan8@uncc.edu

Hunter Burnett
The William States Lee College of Engineering
University of North Carolina at Charlotte
Charlotte, United States of America
hburnet7@uncc.edu

Lauren Bourque
The William States Lee College of Engineering
University of North Carolina at Charlotte
Charlotte, United States of America
lbourqu1@uncc.edu

Michael Morra
The William States Lee College of Engineering
University of North Carolina at Charlotte
Charlotte, United States of America
mmorra@uncc.edu

***Abstract*—This report investigates crime prediction in Charlotte, North Carolina, employing classical and deep learning models on a dataset comprising over 276,000 crime incidents reported by the Charlotte Mecklenberg Police Department from 2016 to 2023. Eight distinct crime classes are considered in the analysis. The classical model, incorporating traditional statistical techniques such as a Naive Bayesian classifier, establishes a baseline for performance evaluation, while the deep learning model implements a Spatiotemporal model to capture spatial and temporal patterns within the crime data. Comparative assessments are conducted, considering accuracy, precision, recall, and F1-score for each crime class. The results offer insights into the strengths and limitations of each approach, contributing valuable perspectives for the causes of various types of crime and their ability to be predicted.**

***Index Terms*—Machine Learning, Deep Learning, Crime, Public Safety**

I. Introduction and Motivation

As cities evolve, so do the challenges in keeping communities safe. In this report, the world of crime prediction in Charlotte, North Carolina is explored, using a dataset spanning 2016 to 2023 from the Charlotte Mecklenberg Police Department. With over 276,000 data points covering 8 different types of crime, the goal was to explore how classical models and cutting-edge deep learning can help society understand and predict crime patterns. Charlotte's unique blend of social and geographical dynamics adds a real-world layer to the problem, highlighting the practical implications of these models in a dynamic urban setting.

The main goal of this project was to explore the common factors causing various types of crime in an urban setting such as location, date, race, education, poverty, and age. This project sought to identify patterns among various types of crime, providing valuable insight to society on how to better mitigate crime.

The link to the GitHub repository can be found here: Click here to view the code

II. Dataset and Training Setup

A. Creating the Dataset

The below datasets were used in creating the dataset for the model:

- **Crime:** https://data.charlottenc.gov/datasets/charlotte::cmpd-incidents-1/about
- **Social Aspects:** https://data.charlottenc.gov/datasets/charlotte::vulnera
to-displacement-by-npa/about
- **2020 Census by Zip Code:** https://www.northcarolina-demographics.com/zip_codes_by_population
- **NIBRS Codes:** https://ucr.fbi.gov/nibrs/2011/resources/nibrs-offense-codes

The first dataset for crime incidents in Charlotte was used as the foundation for the finalized dataset. It contained over 600,000 data points with data on the date, latitude, longitude, zip code, place description, NIBRS code and description, and Neighborhood Profile Area ID (NPA). The NPA ID is a numerical code used to identify neighborhoods in Charlotte. The NIBRS codes are the FBI's system for identifying and classifying a crime and labeling it as a crime against a person, property or society.

This dataset was combined with a dataset containing sociological data by NPA in Charlotte. The data in this dataset included: percent of the population with a high school diploma, percent of non-white population, percent of population living under poverty, percent of population owning a home, along with binary counts if each percentage was above 50%. It also contained a score from 1-5 on the likelihood of displacement and a binary value determining if that NPA was vulnerable to displacement. The NPA from used to match the data from this dataset with the original dataset for crime incidents.

Lastly, the dataset containing the 2020 census by zip code for Charlotte was combined with the original dataset. A column was added called 'Population' where the census data was stored, using the zip code in each crime incident to match census data by zip code.

B. Pre-Processing

The code for cleaning the dataset can be found in the Data Manipulation folder on the GitHub repository.

The dataset was cleaned to ensure the model was trained properly. Any rows in the data set with empty (NaN) values were removed from the dataset to avoid providing incomplete data. The dates the incident was reported and ended were removed from the dataset to only focus on the date the incident began. Incidents that had the address as the location where the officer took the report instead of where the incident occurred were also removed from the dataset to prevent location data from being inaccurate. Unconfirmed incident reports were also removed.

A 'Violent-Crime' column was created in the dataset, where any incident with the 'Crime Against' column value being 'Person' was set to 1. This ruled that any crime happening against a person was considered a violent crime. After this, any variables with string values were mapped to integer values. The enumerations for these integer values can be found in the Enumerations folder on the GitHub repository.

This resulted in a dataset with 24 input features and slightly over 276,000 data points. An 80/20 training and test split was used to train the model.

The overall architecture of the pre-processing can be seen below in Figure 1.

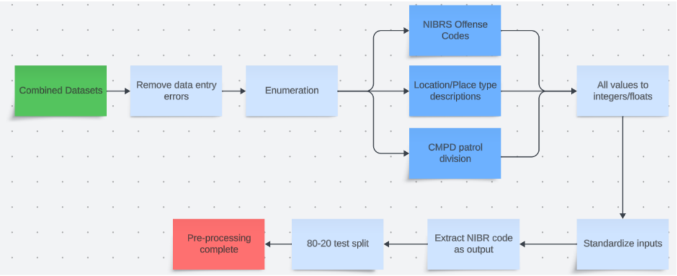


Fig. 1. Pre-processing architecture.

III. Approach

A. Classification Revisions

The goal and scope of the project shifted through various revisions. Various strategies of classification were tested and are listed below.

a) Binary Classification: At first, the model was performing binary classification, classifying crimes as either violent or non-violent. In this case, the 'Violent-Crime' column was used as the ground truth, with a crime against a person being considered violent and a crime against society or property being considered non-violent. This model proved to be accurate, but these results aren't as applicable or useful for society.

b) Expanding Violent Classes: Next, 13 total classes were created, with one class for all non-violent crimes and 12 classes for the various violent crimes. This data skewed the model and made it overfit to non-violent crimes. In this strategy, crimes were only predicted to be non-violent.

c) Classifying Each Crime: The next strategy attempted was classifying each crime. However, there are 45 different types of crime, all with varying numbers of reported crimes. This caused the model to be heavily weighted and much more accurate at predicting only specific crimes.

d) Class Reduction: In the original approach, 45 crime classes were used. However the classification accuracy across under-sampled classes was poor and the model was subjected to overfitting due to uneven data point distribution among classes. This overemphasis on crimes with more data points led to a simplification in the number of classes considered in the model down to 8 distinct classes of crimes. They are as follows: (1) Assault (2) Thefts (3) Vehicle Thefts (4) Property Offenses (5) Sexual Offenses (6) Money Offenses (7) Drug Offenses (8) Falsification. This led to an improvement in classification accuracy in under-sampled classes as well as a more fair distribution of data points across all classes.

B. Various Models

Both classical and deep learning models were trained to explore the effectiveness and accuracy of each type of model, allowing for a better selection of the proper model for this problem. This also allowed for the direct exploration of the results for classical versus deep learning models.

C. Classical Models

Various classical models were implemented and tested to compare the results and discover which model proved most accurate. A Naive Bayesian classifier and Logistic Regression model were implemented, with the Naive Bayesian model yielding the best results.

D. Deep Learning Models

A Spatiotemporal model was implemented for its ability to capture the relationship between variables that can change over space and time. It also allows for the ability to handle data that occurs over irregular intervals in time. This was necessary, as crime is extremely irregular. The various layers are listed below.

a) 1D Convolutional Layer: The model contained a 1D Convolutional layer used to capture spatial patterns in the data using latitude and longitude. A kernel size of 3 was used to identify local patterns in the data. This layer produced 32 outputs.

b) ReLU Activation Function: Other layers in the model are linear. This layer introduced non-linearity to the model, allowing the network to capture complex patterns in the data. A ReLU activation function was chosen because of its efficient training time.

c) LSTM Layer: This layer was used to capture time-based patterns in the data. It was designed to address the vanishing gradient problem where gradients become too small during backwards propagation through many time steps during training. This layer has the ability to capture long-range patterns and dependencies in the data, where most networks can't. A total of 50 outputs were produced from this layer.

d) Fully Connected Layer: This layer was used to combine the learned features from the convolutional and LSTM layers to make a final prediction.

The overall model structure can be seen below in Figure 2.

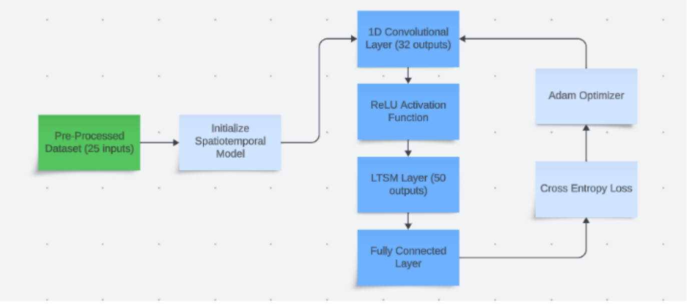
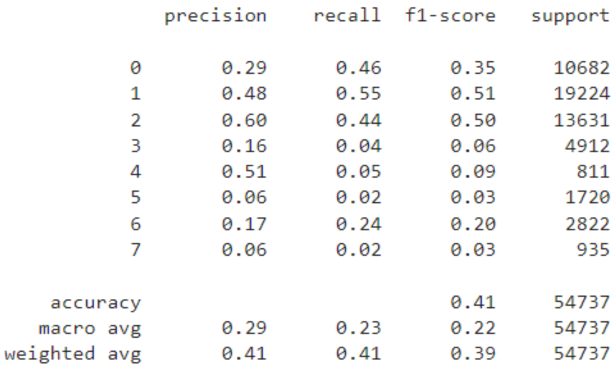


Fig. 2. Deep learning architecture.

IV. Results and Analysis

A. Classical Model Results

The results from the Naive Bayesian classifier can be seen in Figure 3 below.



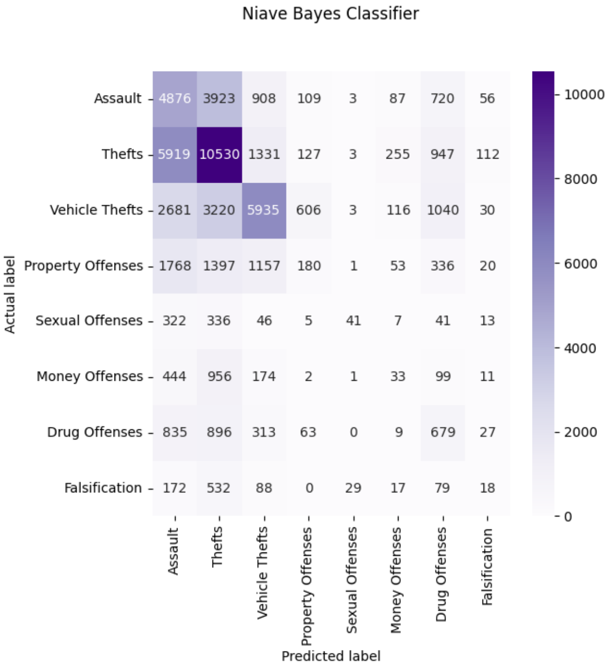
```
              precision    recall  f1-score   support

           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. 3. Naive Bayesian results.

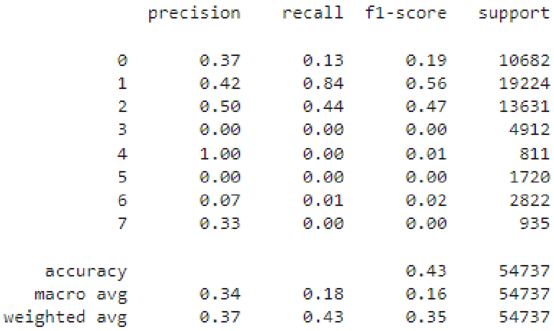
The confusion matrix for the Naive Bayesian classifier can be seen in Figure 4 below.



	Assault	Thefts	Vehicle Thefts	Property Offenses	Sexual Offenses	Money Offenses	Drug Offenses	Falsification
Assault	4876	3923	908	109	3	87	720	56
Thefts	5919	10530	1331	127	3	255	947	112
Vehicle Thefts	2681	3220	5935	606	3	116	1040	30
Property Offenses	1768	1397	1157	180	1	53	336	20
Sexual Offenses	322	336	46	5	41	7	41	13
Money Offenses	444	956	174	2	1	33	99	11
Drug Offenses	835	896	313	63	0	9	679	27
Falsification	172	532	88	0	29	17	79	18

Fig. 4. Naive Bayesian confusion matrix.

The results from the Logistic Regression classifier can be seen in Figure 5 below.



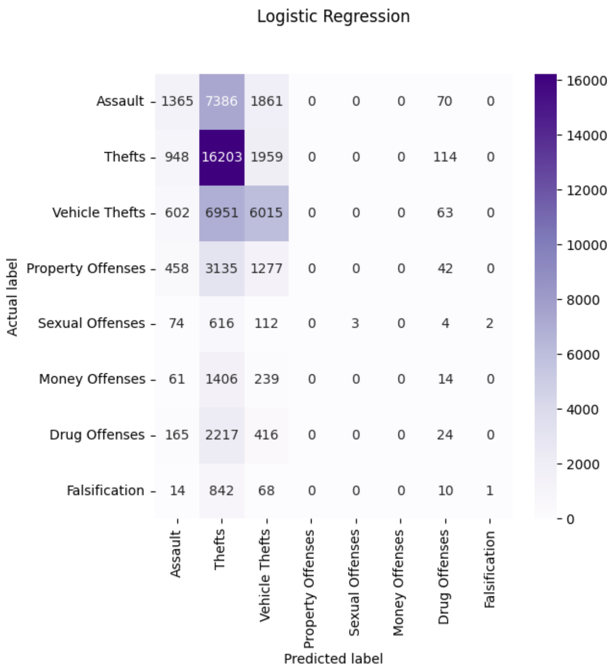
```
              precision    recall  f1-score   support

           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.

The confusion matrix for the Logistic Regression classifier can be seen in Figure 6 below.



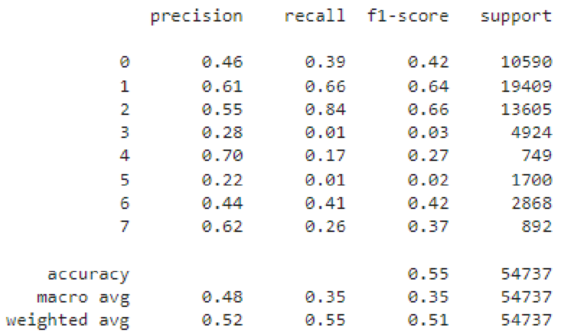
Actual label	Assault	Thefts	Vehicle Thefts	Property Offenses	Sexual Offenses	Money Offenses	Drug Offenses	Falsification
Assault	1365	7386	1861	0	0	0	70	0
Thefts	948	16203	1959	0	0	0	114	0
Vehicle Thefts	602	6951	6015	0	0	0	63	0
Property Offenses	458	3135	1277	0	0	0	42	0
Sexual Offenses	74	616	112	0	3	0	4	2
Money Offenses	61	1406	239	0	0	0	14	0
Drug Offenses	165	2217	416	0	0	0	24	0
Falsification	14	842	68	0	0	0	10	1

Fig. 6. Logistic Regression confusion matrix.

As can be seen from the results, the Logistic Regression classifier provided a better overall accuracy, but the Naive Bayesian classifier provided overall better F1 scores for the classes.

B. Deep Learning Model Results

The results from the Spatiotemporal model can be seen in Figure 7 below.



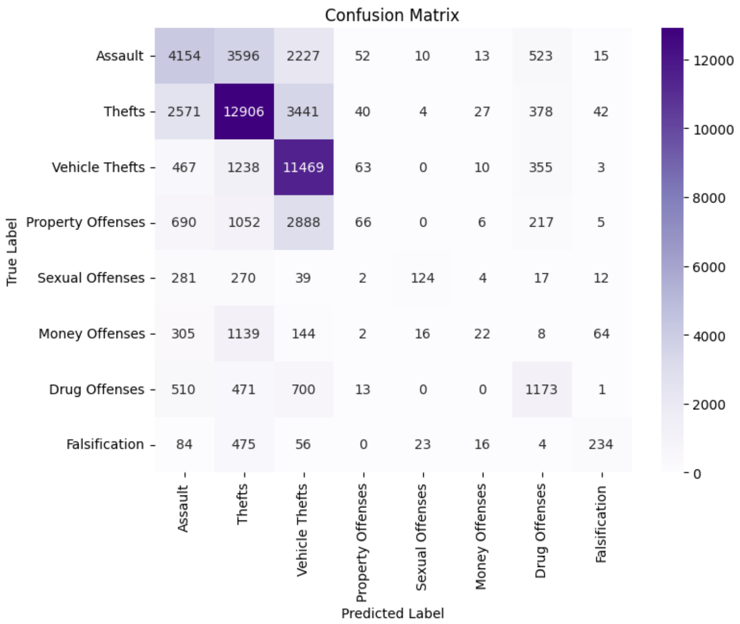
```
              precision    recall  f1-score   support

           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. 7. Spatiotemporal results.

The confusion matrix for the Logistic Regression classifier can be seen in Figure 8 below.



True Label	Assault	Thefts	Vehicle Thefts	Property Offenses	Sexual Offenses	Money Offenses	Drug Offenses	Falsification
Assault	4154	3596	2227	52	10	13	523	15
Thefts	2571	12906	3441	40	4	27	378	42
Vehicle Thefts	467	1238	11469	63	0	10	355	3
Property Offenses	690	1052	2888	66	0	6	217	5
Sexual Offenses	281	270	39	2	124	4	17	12
Money Offenses	305	1139	144	2	16	22	8	64
Drug Offenses	510	471	700	13	0	0	1173	1
Falsification	84	475	56	0	23	16	4	234

Fig. 8. Spatiotemporal confusion matrix.

These results show that the deep learning models are more accurate and provide a better prediction than the classical models. This also shows that the classes with more crime incidents have a better rate of prediction than the classes with less incidents. It also shows that the features provided in the dataset have a much higher impact on the prediction of crimes such as thefts and assault rather than money offenses or drug offenses.

The predicted versus actual thefts can be seen in Figures 9 and 10 below.

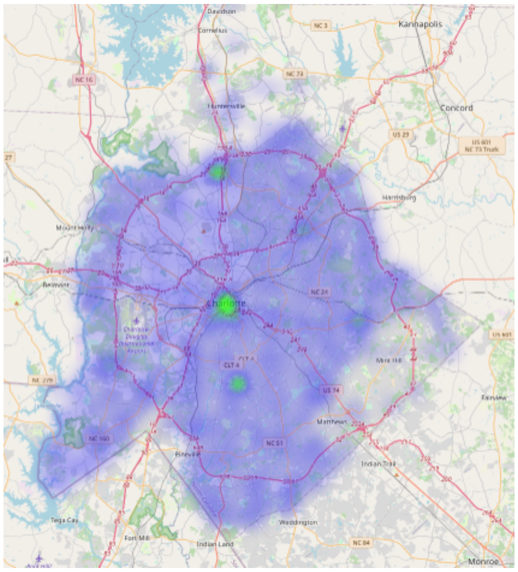


Fig. 9. Predicted theft heatmap.

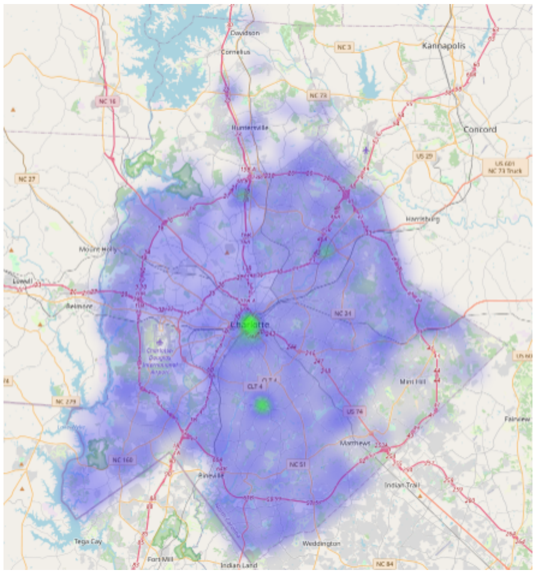


Fig. 10. Actual theft heatmap.

The results show that the predicted thefts are very similar to the actual thefts that occurred.

V. Lessons Learned

The overarching lesson that was learned was that the deep learning model provided much greater accuracy and effectiveness for solving this real world problem as opposed to implementing a classical model. The scope of the project and the strategy to classify data shifted multiple times, demonstrating the importance of flexibility and creative problem-solving in machine learning. Certain classes had a much higher accuracy than others, demonstrating the importance of having a well-balanced dataset. The most time-consuming part of the project's process was creating a meaningful dataset and ensuring that proper pre-processing was done, showing the importance of having clean data in creating an effective model. One feature that impacted the model was race and the fact that it was stored as a binary non-white or white value. This groups all non-white races into the equation, biasing the data and influencing the model to predict more crime in non-white areas. This process demonstrated the importance of feature selection and how one or two features can entirely skew a model. In order to improve the model access to specific demographic information is needed.