CSE 3000 – Final Project Report

John Morgan – jdm21023, Marcus Maravilla – mjm21037, Harvansh Pelia – hsp22002, Jobe Goetz – jjg20006

GitHub Name: cse3000-predictive-crime-modeling, from Jobe Goetz - LINK

1. Abstract

This project develops a predictive model to classify the type of crime based on complaint data provided by the NYPD. The dataset includes all valid crimes reported to NYPD, where each row is a complaint. The most up to date dataset can be found here. Using a Random Forest Classifier, we trained a multiclass and binary classification model on time-based and location-based features. We conclude with ethical reflections and propose mitigation strategies.

2. Introduction

Predictive policing has gained attention as a tool for law enforcement agencies aiming to allocate resources more efficiently. However, predictive models risk reinforcing historical biases if not carefully analyzed. This project focuses on developing a model that predicts the type of crime and occurrence of crime based on complaint data, while critically examining socioeconomic biases within the predictions.

3. Literature review

   - Predictive Policing Explained
   - Artificial Intelligence in Predictive Policing Issue Brief
   - Algorithm predicts crime a week in advance, but reveals bias in police response

4. Methods

   - We used the **NYPD Complaint Data (Year-to-Date)** dataset, selecting the following attributes:
     - Latitude – latitude of the crime
     - Longitude – longitude of the crime
     - CMPLNT_FR_DT - Exact date of occurrence for the reported event
     - CMPLNT_FR_TM - Exact time of occurrence for the reported event
   - Features:
     - Month (extracted from CMPLNT_FR_DT)

    - DayOfWeek (extracted from CMPLNT_FR_DT)
    - Hour (extracted from CMPLNT_FR_TM)
    - Latitude
    - Longitude

- We used a **Random Forest Classifier** for multiclass classification, predicting the type of crime and occurrence of crime.
- Model performance was assessed using back testing.
- Model bias was performed by analyzing the median incomes of boroughs.
- Implementation

  Libraries used: pandas, folium, scikit-learn, numpy

    - **pandas –** data manipulation
    - **scikit-learn** – machine learning model building and evaluation
    - **folium** – interactive maps
    - **numpy** – numerical operations

- Our implementation can be broken down into the following steps (see corresponding comments in our code).

1. Load and preprocess data

   Loads and cleans up the NYPD complaint data. Here, we preprocess the data by dropping irrelevant columns, removing rows with missing values, and extract the time and location-based features for our model.

2. Prepare Data

   For binary classification, we need to generate locations with no crime in our data to more accurately train and test our model (crime will have always occurred in the complaint data). We do this by randomly generating different locations that have "No Crime" and merge this data into our original dataset. Finally, we split the dataset into training (70%) and testing (30%) sets.
   For multiclass classification, it's more straightforward. We prepare our features and split the dataset into training (70%) and testing (30%) sets.

3. Train and Evaluate Models

   This is where we train and evaluate the 2 Random Forest Prediction models (binary for crime occurrence and multiclass classification for type of crime). The accuracy score and classification report is then produced using scikit-learn.

4. Mapping

   This is where we map the crime occurrence and crime type predictions vs actual. For crime occurrences, green markers show correct predictions, while red markers show incorrect ones. For crime type, each marker shows the predicted crime type vs the actual crime type.

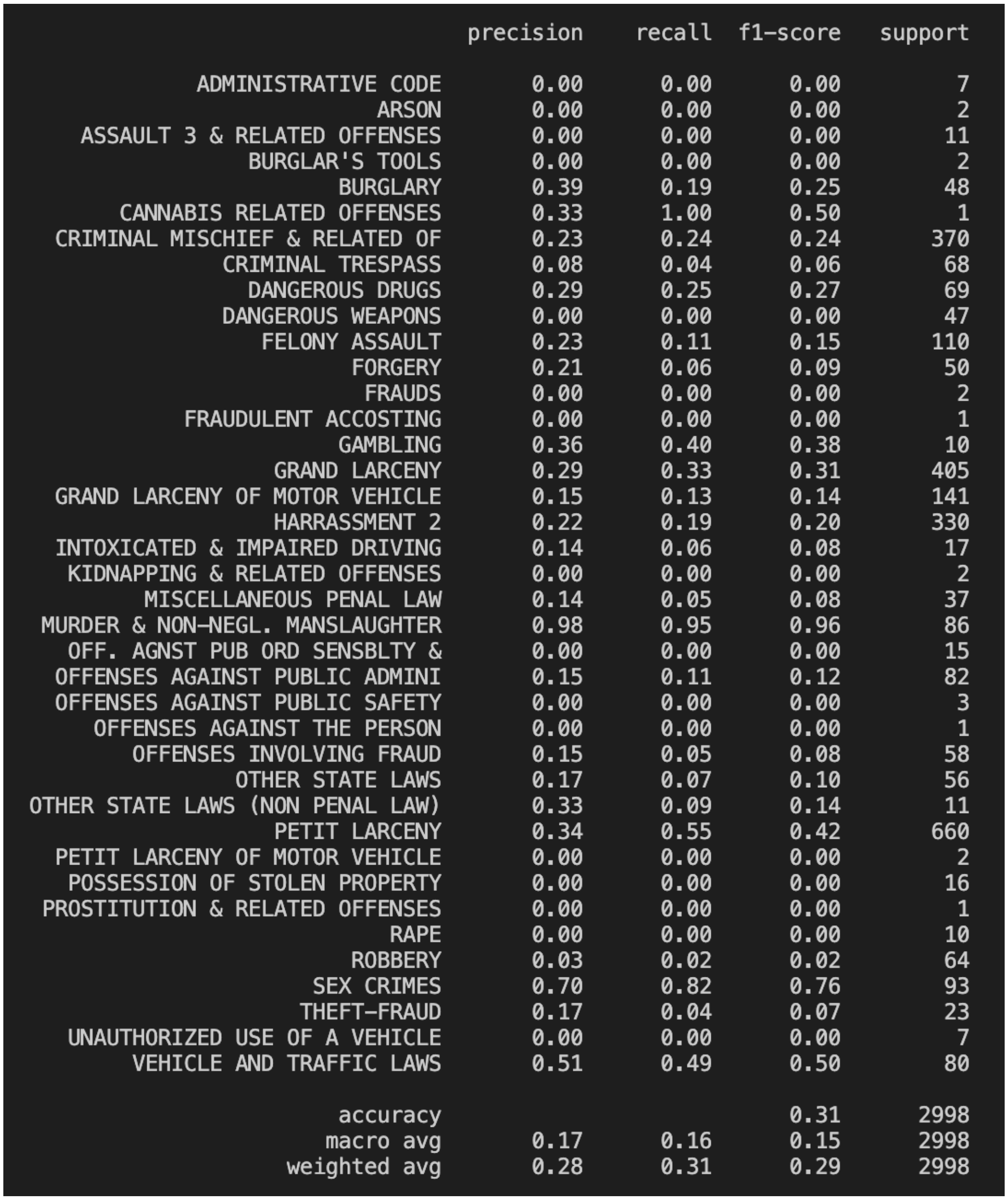
5. Back testing

   Back testing is done monthly. For each month, the model is trained on all the previous month's data and makes predictions for the next month.

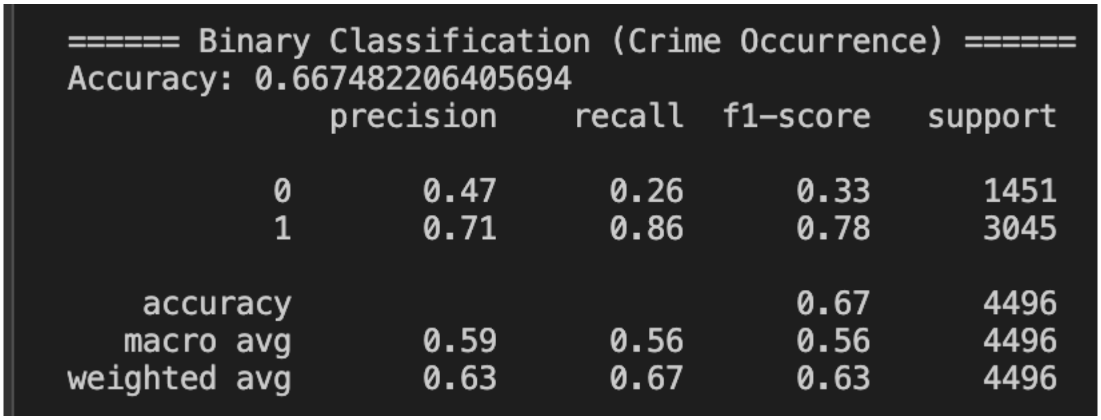
6. Bias testing

The New York Census dataset is used to add income information for each borough. The data is then split into low- and high-income groups based on median income, ran through the model, then evaluated to see if there's any bias.

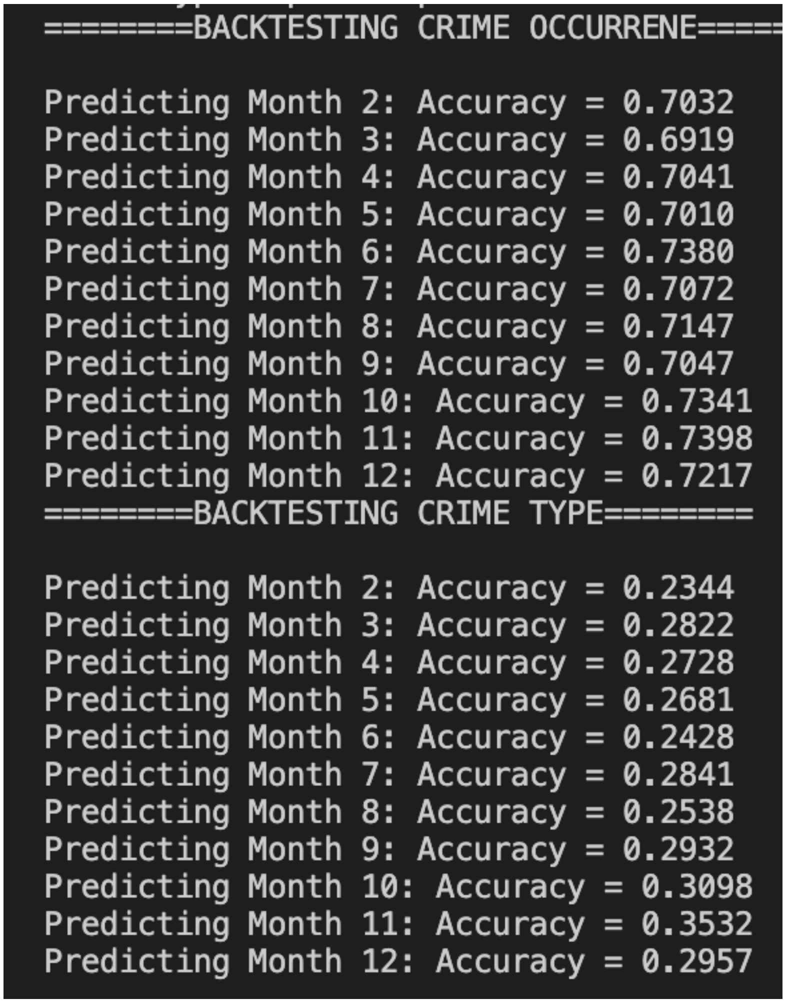
5. Results



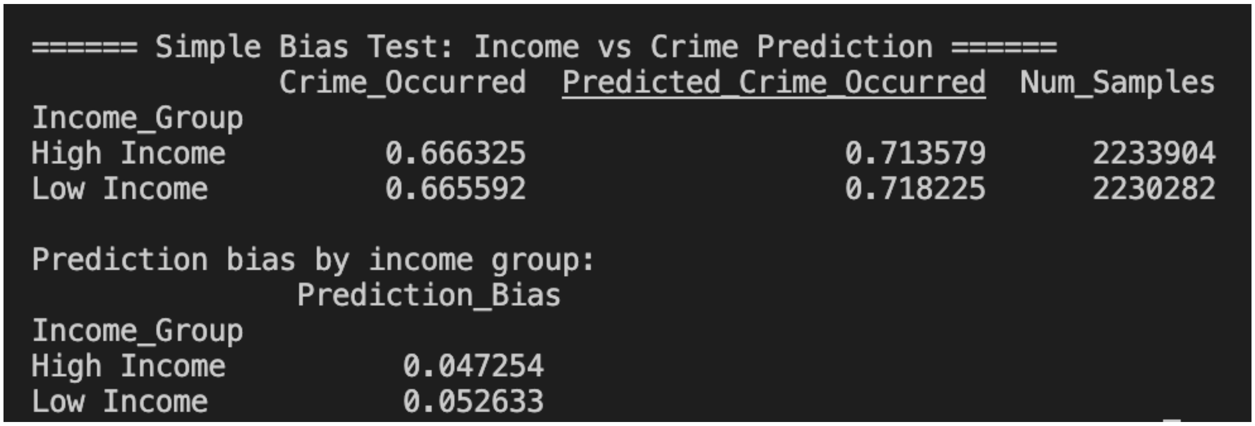
```
                                  precision    recall  f1-score   support

             ADMINISTRATIVE CODE       0.00      0.00      0.00         7
                           ARSON       0.00      0.00      0.00         2
    ASSAULT 3 & RELATED OFFENSES       0.00      0.00      0.00        11
                 BURGLAR'S TOOLS       0.00      0.00      0.00         2
                        BURGLARY       0.39      0.19      0.25        48
       CANNABIS RELATED OFFENSES       0.33      1.00      0.50         1
  CRIMINAL MISCHIEF & RELATED OF       0.23      0.24      0.24       370
               CRIMINAL TRESPASS       0.08      0.04      0.06        68
                 DANGEROUS DRUGS       0.29      0.25      0.27        69
               DANGEROUS WEAPONS       0.00      0.00      0.00        47
                  FELONY ASSAULT       0.23      0.11      0.15       110
                         FORGERY       0.21      0.06      0.09        50
                          FRAUDS       0.00      0.00      0.00         2
            FRAUDULENT ACCOSTING       0.00      0.00      0.00         1
                        GAMBLING       0.36      0.40      0.38        10
                   GRAND LARCENY       0.29      0.33      0.31       405
  GRAND LARCENY OF MOTOR VEHICLE       0.15      0.13      0.14       141
                   HARRASSMENT 2       0.22      0.19      0.20       330
  INTOXICATED & IMPAIRED DRIVING       0.14      0.06      0.08        17
   KIDNAPPING & RELATED OFFENSES       0.00      0.00      0.00         2
         MISCELLANEOUS PENAL LAW       0.14      0.05      0.08        37
 MURDER & NON-NEGL. MANSLAUGHTER       0.98      0.95      0.96        86
   OFF. AGNST PUB ORD SENSBLTY &       0.00      0.00      0.00        15
  OFFENSES AGAINST PUBLIC ADMINI       0.15      0.11      0.12        82
  OFFENSES AGAINST PUBLIC SAFETY       0.00      0.00      0.00         3
     OFFENSES AGAINST THE PERSON       0.00      0.00      0.00         1
        OFFENSES INVOLVING FRAUD       0.15      0.05      0.08        58
                OTHER STATE LAWS       0.17      0.07      0.10        56
OTHER STATE LAWS (NON PENAL LAW)       0.33      0.09      0.14        11
                   PETIT LARCENY       0.34      0.55      0.42       660
  PETIT LARCENY OF MOTOR VEHICLE       0.00      0.00      0.00         2
   POSSESSION OF STOLEN PROPERTY       0.00      0.00      0.00        16
 PROSTITUTION & RELATED OFFENSES       0.00      0.00      0.00         1
                            RAPE       0.00      0.00      0.00        10
                         ROBBERY       0.03      0.02      0.02        64
                      SEX CRIMES       0.70      0.82      0.76        93
                     THEFT-FRAUD       0.17      0.04      0.07        23
   UNAUTHORIZED USE OF A VEHICLE       0.00      0.00      0.00         7
        VEHICLE AND TRAFFIC LAWS       0.51      0.49      0.50        80

                        accuracy                           0.31      2998
                       macro avg       0.17      0.16      0.15      2998
                    weighted avg       0.28      0.31      0.29      2998
```



```
====== Binary Classification (Crime Occurrence) ======
Accuracy: 0.667482206405694
              precision    recall  f1-score   support

           0       0.47      0.26      0.33      1451
           1       0.71      0.86      0.78      3045

    accuracy                           0.67      4496
   macro avg       0.59      0.56      0.56      4496
weighted avg       0.63      0.67      0.63      4496
```



```
========BACKTESTING CRIME OCCURRENE=====

Predicting Month 2: Accuracy = 0.7032
Predicting Month 3: Accuracy = 0.6919
Predicting Month 4: Accuracy = 0.7041
Predicting Month 5: Accuracy = 0.7010
Predicting Month 6: Accuracy = 0.7380
Predicting Month 7: Accuracy = 0.7072
Predicting Month 8: Accuracy = 0.7147
Predicting Month 9: Accuracy = 0.7047
Predicting Month 10: Accuracy = 0.7341
Predicting Month 11: Accuracy = 0.7398
Predicting Month 12: Accuracy = 0.7217
========BACKTESTING CRIME TYPE========

Predicting Month 2: Accuracy = 0.2344
Predicting Month 3: Accuracy = 0.2822
Predicting Month 4: Accuracy = 0.2728
Predicting Month 5: Accuracy = 0.2681
Predicting Month 6: Accuracy = 0.2428
Predicting Month 7: Accuracy = 0.2841
Predicting Month 8: Accuracy = 0.2538
Predicting Month 9: Accuracy = 0.2932
Predicting Month 10: Accuracy = 0.3098
Predicting Month 11: Accuracy = 0.3532
Predicting Month 12: Accuracy = 0.2957
```



```
====== Simple Bias Test: Income vs Crime Prediction ======
              Crime_Occurred  Predicted_Crime_Occurred  Num_Samples
Income_Group
High Income         0.666325                  0.713579      2233904
Low Income          0.665592                  0.718225      2230282

Prediction bias by income group:
               Prediction_Bias
Income_Group
High Income           0.047254
Low Income            0.052633
```

- Different accuracy for different types of crimes:
    - Very good at predicting murders and ok at predicting crimes like Petit Larceny
    - Not so good at predicting crimes that are rarer such as arson
- Bias Analysis:
    - There is negligible bias because our model isn't trained on income
    - However, this does not guarantee that the data from the NYPD isn't biased

6. Discussion
    - The Random Forest model performs well overall, but demographic discrepancies suggest that the model may be influenced by underlying biases in historical policing practices.
    - Predictive crime models risk reinforcing racial disparities if not carefully audited.
    - Arrest data may not reflect true crime rates but rather policing patterns.
    - Mitigation strategies include:
        - Balancing training data.
        - Incorporating socioeconomic factors rather than demographics alone.
        - Transparent reporting of model limitations.

7. Conclusion

This project demonstrates that Random Forest models can effectively predict crime occurrence (binary) and crime type (multiclass) using NYPD complaint data. The binary model performs well overall, but the multiclass model struggles with rare crime types like arson. While income-based bias was minimal, potential bias in the original data remains a concern.

To improve the binary model, we could enhance the "no crime" data by simulating more realistic negative samples based on population density or time of day. For the multiclass model, better performance could be achieved by addressing class imbalance through oversampling or class weighting, and by adding more contextual features like neighborhood type or historical crime trends. Testing different algorithms such as Gradient Boosting or neural networks may also yield improved accuracy.

Overall, predictive policing tools must be developed carefully to avoid reinforcing bias and should be continuously refined with broader data sources and community input.

8. Reference list

- Maheshwari, M., & Dahiya, K. (2024). *Crime prediction model using three classification techniques*. International Journal of Advanced Computer Science and Applications (IJACSA), 15(1). https://doi.org/10.14569/IJACSA.2024.0150123
- New York City Police Department (NYPD). (2024). *NYPD Arrest Data (Year-to-Date)*. NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc

- NYPD Arrest Data (Year-to-Date). New York City Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc
- Lum, K., & Isaac, W. (2016). *To predict and serve?* Significance, 13(5), 14–19. https://arxiv.org/abs/1610.01943
- New York City Police Department (NYPD). (2024). *NYPD Arrest Data (Year-to-Date)*. NYC Open Data. https://data.cityofnewyork.us/Public-Safety/NYPD-Arrest-Data-Year-to-Date-/uip8-fykc
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org. Retrieved from https://fairmlbook.org/