DATA 606

Capstone in Data Science (06.7464) SP2023

Final submission:

Baltimore City crime rate and safety level statistics.

- Singireddy Karthik Reddy
TP48091.

Introduction:

1. Many causes are contributing to the rise in crime in today's globe. If we look closely, most crimes of a particular type tend to occur in the same locality, possibly due to the individuals that live there.
2. We are working on a model that assesses which areas are prone to specific types of crimes, allowing us to assess the area's safety. This will benefit the US Department of Homeland Security. Also, for those who are looking to start a new life in the city.

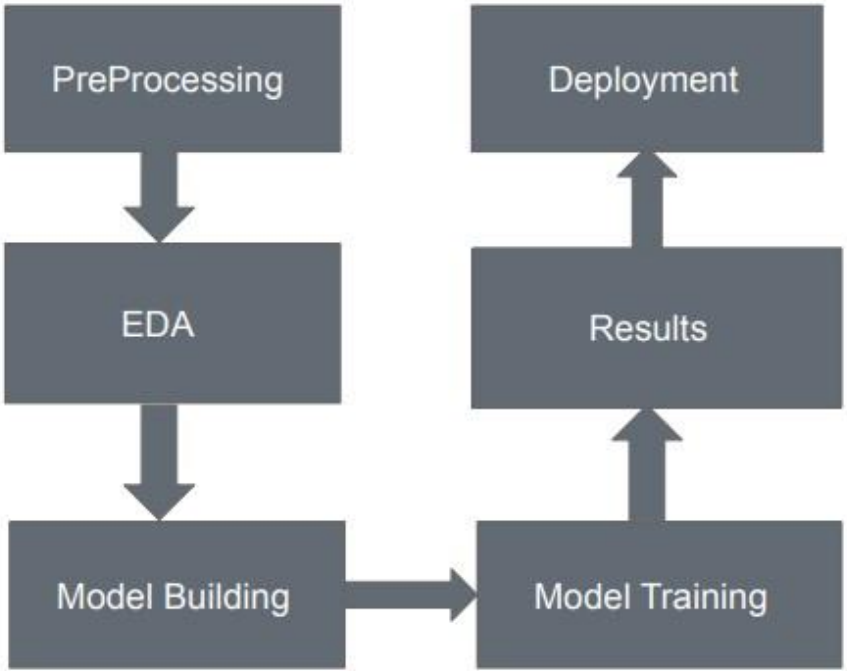
Dataset:

- We are using a dataset from the Open Baltimore database of the United States Government for the city of Baltimore.
- The dataset contains crime details like crime code, description, location of the event, date and time of report, etc.
- The dataset provides information about all the crimes that were reported in the city of Baltimore.

Description of the dataset:

Variable Name	Description
RowID	Serial Number
CrimeDateTime	Date and time of the criminal incident
CrimeCode	Unique code of crime by crime dept.
Location	Location of the incident
Description	Description/Type of the Incident
Inside_Outside	If the attack was inside or outside
Weapon	Weapon used for the attack
Post	Postal Code
District	District of Incident.
Neighborhood	Neighborhood of the incident.
Latitude	Latitude of location of incident.
Longitude	Longitude of location of incident.
GeoLocation	Geo Location of the incident.
Premise	Premise of the incident.
VRIName	Video remote interpretation.
Total_Incidents	Total incidents in the location

Project Framework:



Preprocessing:

1. Removed null values.
2. Dropped unwanted columns like location and latitude.
3. Target column: Level (obtained from the description).
4. Indexed the target column using a string indexer.

- Low : 0 (Larceny, Burglary, Auto Theft).
- Medium: 1 (Robbery, Assault).
- High: 2 (Shooting, Rape, Arson, Homicide).

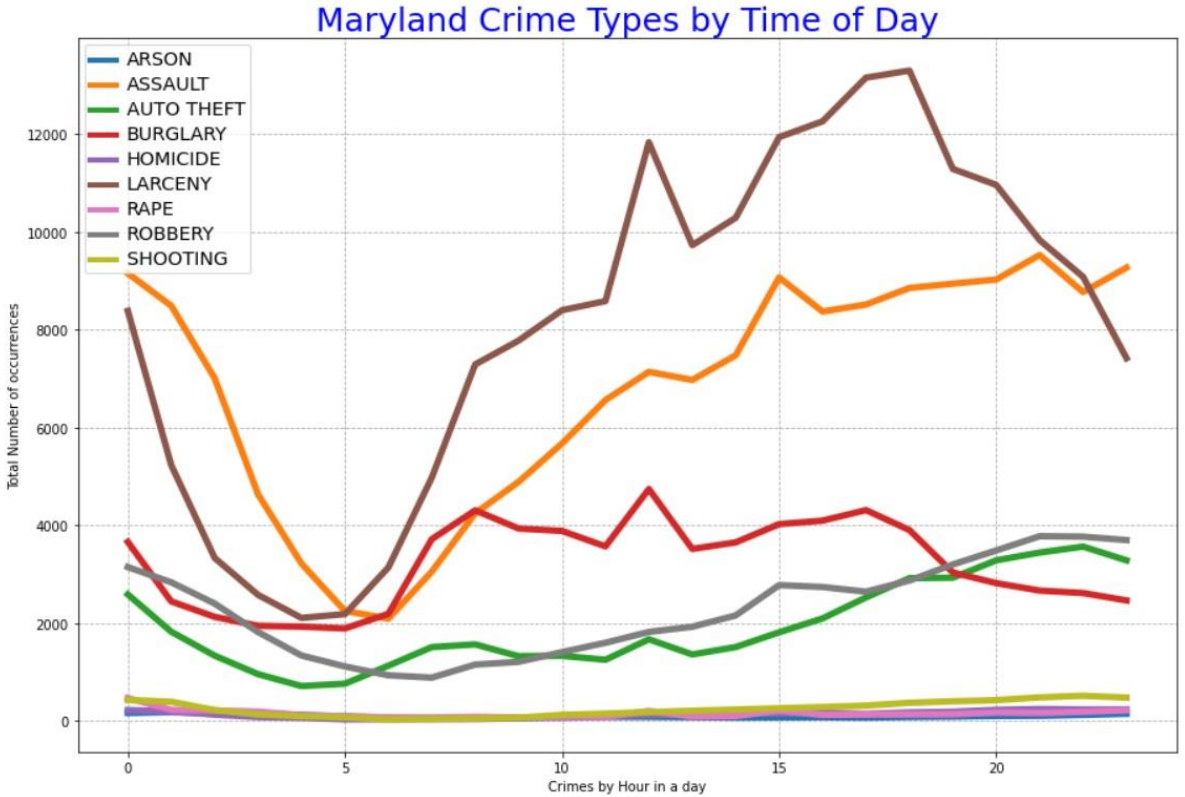
Libraries used:

- Matplotlib.
- Numpy.
- StreamLit.
- Pandas.

Exploratory Data Analysis:

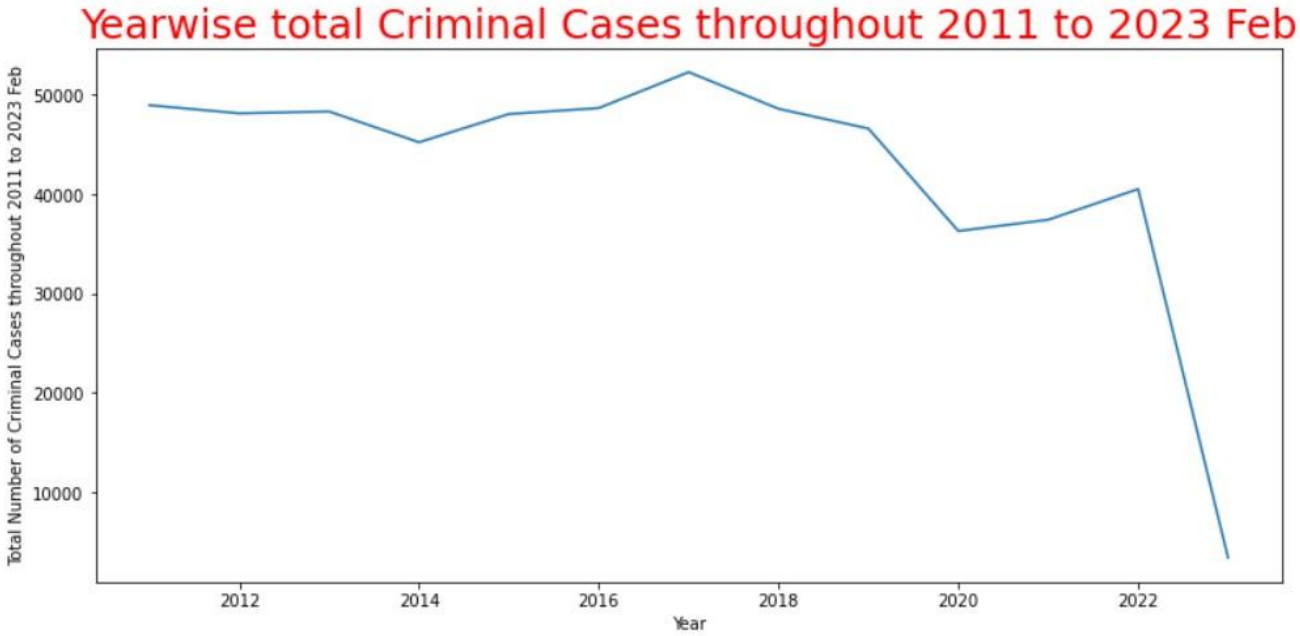
Hypothesis:

- Line Plot of occurrences of crimes through the hour of the day.
- From this graph, we can see that Crime is high from 10:00 AM to 18:00 PM.



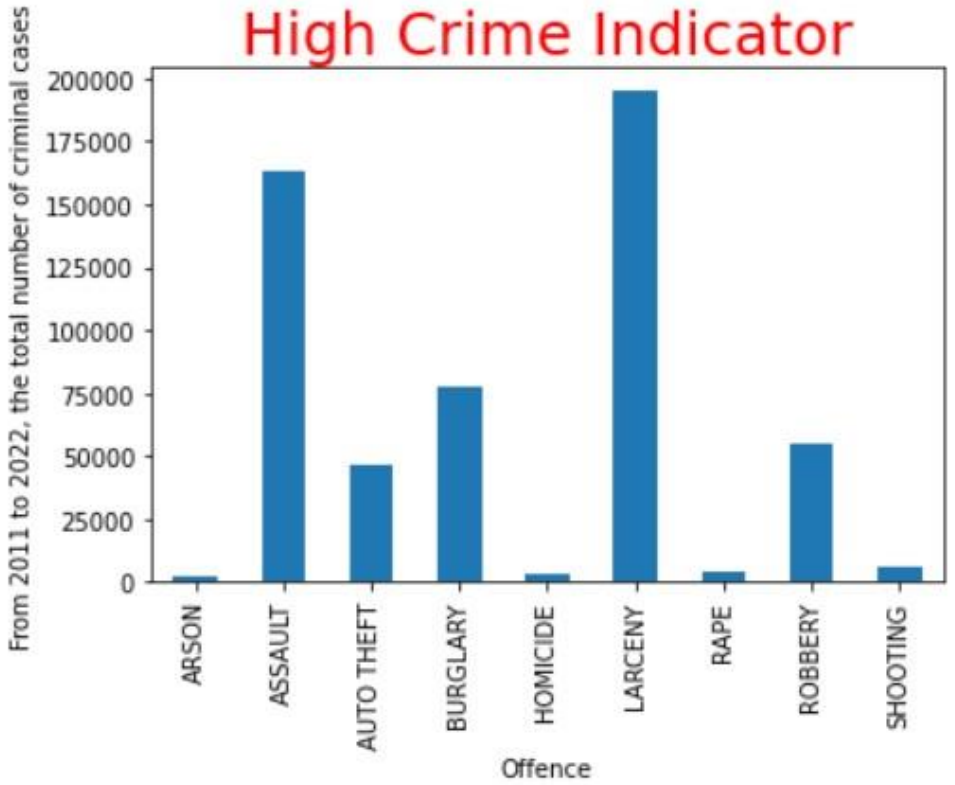
Hypothesis:

- We see a decrease in the number of attacks after the covid-19 Pandemic.



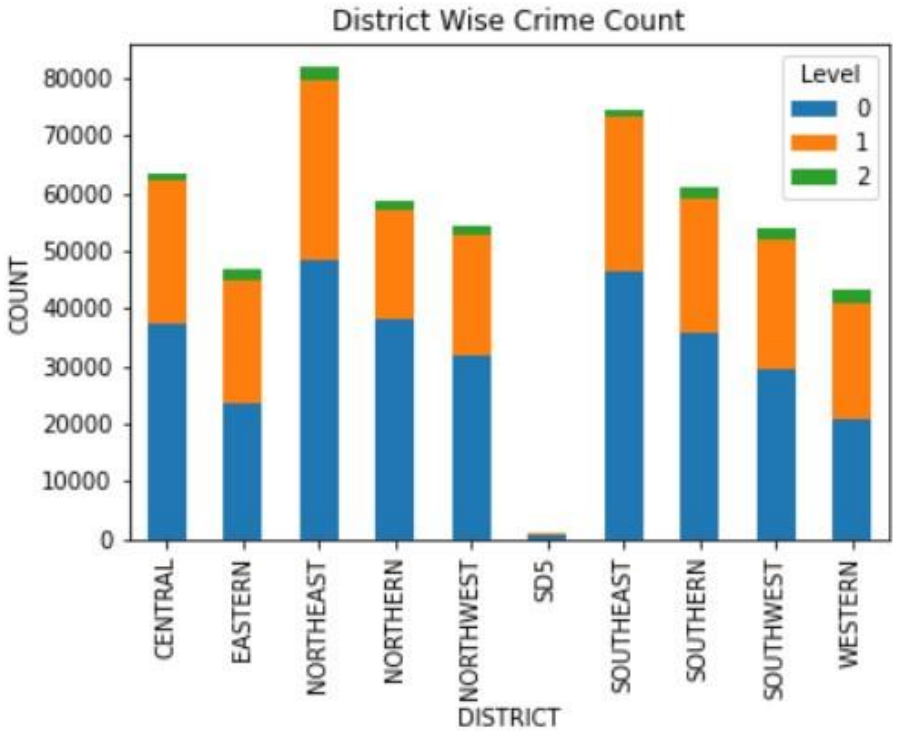
Hypothesis:

- We see the highest cases were reported for Larceny, followed by Assault, burglary, Robber, etc.



Hypothesis:

- We see the highest cases were reported in the Northeast district followed by the Southeast.



Implementing Machine learning models:

To develop machine learning models, we implemented 2 approaches:

- 2-way classification.
- 3-way classification.

Random Forest classification:

- The random forest consists of a large number of individual decision trees that operate as an ensemble.
- Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

Random Forest Classification (2-way Classification)

```
Accuracy of Random Forest with Balanced class weight:  0.962461817301157
[[122610   1776]
 [  3029    588]]
              precision    recall  f1-score   support

   NON_FATAL       0.98      0.99      0.98    124386
       FATAL       0.25      0.16      0.20      3617

    accuracy                           0.96    128003
   macro avg       0.61      0.57      0.59    128003
weighted avg       0.96      0.96      0.96    128003
```

Random Forest Results (3-way Classification)

```
Accuracy of Random Forest with Balanced class weight:  0.5885643305235033
[[21376 27149  1385]
 [20487 53358   631]
 [ 1433  1580   604]]
              precision    recall  f1-score   support

           1       0.49      0.43      0.46     49910
           0       0.65      0.72      0.68     74476
           2       0.23      0.17      0.19      3617

    accuracy                           0.59    128003
   macro avg       0.46      0.44      0.44    128003
weighted avg       0.58      0.59      0.58    128003
```

K-Nearest Neighbor:

- The KNN algorithm assumes that similar things exist in proximity.
- K-nearest neighbor or K-NN algorithm basically creates an imaginary boundary to classify the data.
- When new data points come in, the algorithm will try to predict that to the nearest boundary line.

KNN Results
(3-way Classification) →

	precision	recall	f1-score	support
0	0.61	0.83	0.70	50248
1	0.49	0.26	0.34	34054
2	0.20	0.02	0.04	2607
accuracy			0.58	86909
macro avg	0.43	0.37	0.36	86909
weighted avg	0.55	0.58	0.54	86909

Future work:

As part of Future work, In the User Interaction interface (StreamLit) we are planning to introduce the ‘drop the pin on the map’ option to fetch the address of the location the user is planning to visit. In the future, we try to acquire more data and build a model which can classify the different types of crime.

References

- Alkesh Bharati1, Dr. Sarvanaguru RA.K2," Crime Prediction and Analysis Using Machine Learning", International Research Journal of Engineering and Technology (IRJET).

- O. Llaha, "Crime Analysis and Prediction using Machine Learning," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 496-501.

- W. Safat, S. Asghar, and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," in IEEE Access, vol. 9, pp. 70080-70094, 2021, doi: 10.1109/ACCESS.2021.3078117.

- https://data.baltimorecity.gov/datasets/part1-crime-data/explore?location=39.304390%2C-76.624118%2C10.97&showTable=true

UMBC

Thank you!