Executive Summary: Baltimore Crime Analysis and Prediction

Business Problem

The city of Baltimore continues to face significant challenges related to crime, underscoring the need for effective prevention strategies and resource allocation. This project aims to predict crime hotspots and analyze crime trends to support local authorities in making informed, data-driven decisions. By leveraging historical crime data, the project identifies spatial and temporal patterns and delivers actionable insights to enhance community safety.

Approach

This project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, a comprehensive methodology for data analysis and predictive modeling. Key steps include:

1. Business Understanding:

- Defined objectives to predict high-crime areas and analyze trends.

- Identified project' needs, focusing on actionable insights for law enforcement and city planners.

2. Data Understanding and Preparation:

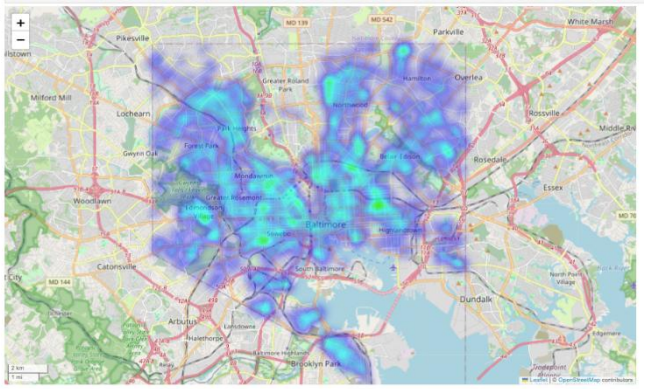
- Utilized historical crime data from Baltimore's open data portal, containing over 25,000 records.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 638033 entries, 0 to 638032
Data columns (total 23 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   RowID            638033 non-null  int64
 1   CCNumber         638033 non-null  object
 2   CrimeDateTime    638033 non-null  object
 3   CrimeCode        638033 non-null  object
 4   Description      638033 non-null  object
 5   Inside_Outside   37723 non-null   object
 6   Weapon           166723 non-null  object
 7   Post             629894 non-null  object
 8   Gender           536109 non-null  object
 9   Age              515351 non-null  float64
 10  Race             607177 non-null  object
 11  Ethnicity        110769 non-null  object
 12  Location         634312 non-null  object
 13  Old_District     566413 non-null  object
 14  New_District     63563 non-null   object
 15  Neighborhood     629097 non-null  object
 16  Latitude         636700 non-null  float64
 17  Longitude        636700 non-null  float64
 18  GeoLocation      638033 non-null  object
 19  PremiseType      586168 non-null  object
 20  Total_Incidents  638033 non-null  int64
 21  x                636700 non-null  float64
 22  y                636700 non-null  float64
dtypes: float64(5), int64(2), object(16)
memory usage: 112.0+ MB
```

- Conducted data preprocessing by cleaning columns with over 70% missing values and standardizing formats for analysis.

3. Modeling and Analysis:

- Performed geospatial analysis using heatmaps to visualize high-risk zones.



- Explored classification models, including **Logistic Regression**, **Random Forest**, and **XGBoost**, to predict crime risk based on spatial and temporal features.

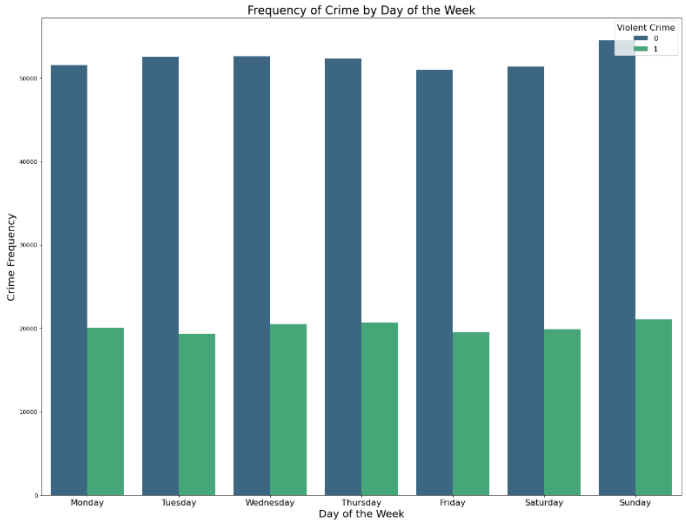
4. Evaluation:

- Assessed model performance using metrics such as accuracy, F1 score, and Area Under the Curve (AUC).

	Model_name	Accuracy_Score	Precision_Score	Recall_Score	F1 Score	ROC AUC Score
6	Xgboost	0.754781	0.520267	0.087071	0.149176	0.530375
5	Gradient Boosting	0.753139	0.800000	0.000179	0.000357	0.500082
0	Logistic Regression	0.753111	1.000000	0.000022	0.000045	0.500011
4	Ada boost	0.753106	0.000000	0.000000	0.000000	0.500000
3	Random Forest Classifier	0.749101	0.460842	0.095461	0.158160	0.529424
1	KNN	0.674042	0.282646	0.208216	0.239788	0.517486
2	Decision Tree	0.630821	0.289564	0.340764	0.313084	0.533338

Key Insights

- Crime Trends: Analysis revealed notable spatial and temporal patterns, with specific neighborhoods experiencing higher crime rates at certain times of the day.



- Hotspot Mapping: Geospatial visualizations effectively identified high-risk zones, facilitating targeted resource deployment.

- Model Performance: Initial modeling efforts yielded promising results, with Xgboost demonstrating superior performance in predicting high-crime areas.

```
Tuned Accuracy: 0.8937
Tuned Precision: 0.9702
Tuned Recall: 0.8124
Tuned F1 Score: 0.8843
Tuned ROC AUC: 0.8937
Adjusted Accuracy: 0.8937
Adjusted Precision: 0.9702
Adjusted Recall: 0.8124
Adjusted F1 Score: 0.8843
Adjusted ROC AUC: 0.8937
```

Evaluation Metrics of Tuned Xgboost

Conclusions and Recommendations

This project illustrates the potential of data-driven approaches to enhance crime prevention efforts in Baltimore. Key recommendations include:

- Using predictive models to allocate resources more efficiently in identified hotspots.

- Integrating temporal data into strategic decision-making to address time-sensitive crime patterns.

- Regularly updating the model with new data to maintain accuracy and relevance.

This analysis underscores the importance of combining geospatial and predictive modeling techniques to tackle urban safety challenges. By adopting these insights, Baltimore can proactively enhance crime prevention measures, improving safety and quality of life for its residents.

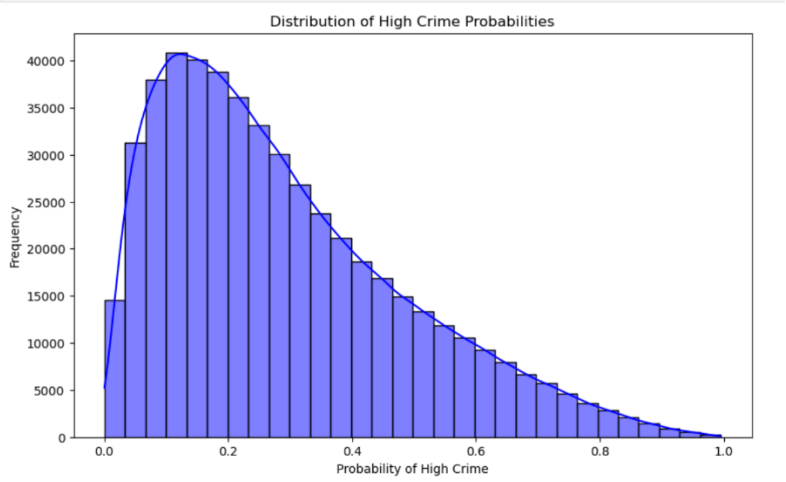


Fig: Predicted Outcome of Xgboost