UK Crime Activity Forecasting

Abstract—**This project involves the study and forecasting of crime trends using open-access, detailed data acquired from the UK Police Open Data portal. A key goal is to construct a Logistic Regression model which is able to predict and classify timeframes and places where crime is above the average. By summarizing crimes by type and month to create a binary classification (High Crime vs. Low Crime), the model is able to achieve a high level of accuracy. The aim is to highlight significant temporal and categorical trends while providing law enforcement and public safety initiatives proactive assistance.**

Keywords—UK Police Data, Crime Analysis, Logistic Regression, Predictive Modeling, Binary Classification

I. Introduction

Analyzing trends and forecasting occurrences of crime is a goal of this project, and I use data available from the UK Police data portal to achieve that goal [1]. I access comprehensive data regarding reported crimes across both England and Wales, and this includes details of the offenses, as well as spatial and temporal data pertaining to the crimes during the months reported.

Criminal behavior is the focus. I aim to construct a model that predicts and defines specific, monthly and spatial parameters when and where offenses will occur. I will analyze the variables along the dimensions of crime, and the months to identify the principal trends and dominant crimes in specific neighborhoods, during the time intervals of focus [3].

The findings will be in the form of analytical dashboards built on Power BI. I intend to present the insights pertaining to spatial and temporal crime trends, clustering tendencies, and overall monthly variation. These insights generated from the descriptive data-analytical model will be of added value to the crime data and will enhance the law enforcement agencies and safety strategists' overall planning and operational efficiency [4].

II. Literature Review

Analyzing the open data initiatives, one can better understand contemporary criminology, including the UK Police Open Data portal [1]. This Open Data portal has made it possible for researchers to move beyond aggregated reporting to fine-grained, street-level analysis. Police records, detailing crime type, location, and date, permit researchers to conduct the critical spatial and temporal analyses needed to identify and describe geographic and temporal 'hot spots' and 'hot times' within which illegal activity is concentrated [5].

Predictive crime modeling aims to foresee crime occurrences in the future. This is commonly framed as a binary problem in crime analytics predicting a given time or place will be "High Crime" or "Low Crime." Successfully converting crime counts into a binary target involves establishing a clear threshold, often the historical mean, a critical step in this analysis. I used the Logistic Regression model as it is one of the most basic statistical models used in the social sciences [6].

This analytical approach aims to generate actionable insights for stakeholders in law enforcement. Although more intricate machine learning models may achieve slight improvements in accuracy, the clear-cut nature of Logistic Regression ends the debate about its suitability in modeling for allocation of discretionary resources, allowing officers and decision-makers to comprehend the elements leading to predictions for prospective high-risk periods. This project uses freely available crime data and a dependable, easy-to-understand model to help initiate data-informed policing, providing a formalized framework for proactive criminal intervention and efficient police resource allocation [7].

III. Methodology

This study followed the Data Analytical Life Cycle (DALC) using a Logistic Regression model for binary classification.

A. Discover

The dataset refers to street-level crime data recorded on the UK Police Data Portal for the area of Central London from January to March 2024 [1]. During the initial phases of data cleaning, I removed rows that had missing critical data, also saving only the columns that I needed for the feature generation process. The two categorical variables, category and month, then needed to be converted to usable numerical features for the predictive model, which was achieved through Label Encoding.

B. Data Preparation

I first aggregated the raw crime data on a monthly basis and categorically, to determine the value of crime_count for each category. The mean of these aggregated monthly crime values then became a threshold for the classification of the binary target variable high_crime (y), which was assigned a value of 1 if the crime count was above the mean and 0 if it was below the mean [6].

C. Plan Model

Logistic Regression was selected given the model's capability for binary classification as well as its interpretability, which is key for actionable insights for law enforcement stakeholders [7]. Its linear configuration makes the understanding of the impact of each feature easier.

D. Build Model

Encoded features (X) as well as target (y) were divided while preserving class proportion into a 75% training and 25% test set. The model was then trained on the training data, and for extensive analysis, the test set was used to derive the predictions (ypred) and probabilities (yprob) [6].

E. Communicate

I explained the model performance by using metrics such as Accuracy and the Classification Report (Precision, Recall, F1-

score). To aid understanding, I directly addressed the model's predictive capacity using the Confusion Matrix and the Receiver Operating Characteristic (ROC) Curve and AUC score [8]. These were aimed at a non-technical audience. I also used a Power BI dashboard for the initial data exploration and trend analysis, presenting the original dataset's crime data geographically and augmented with charts showing crime frequency by category, changes over months, and geographically.

IV. Results and Discussion

The Logistic Regression model was utilized to classify combinations of crime category and month into High Crime or Low Crime activity. The model's predictive performance on the unseen test set demonstrated a strong and reliable classification ability.

Model Evaluation Metrics

The model's performance on the test set was measured using key classification metrics

Accuracy: 0.7738(77.38%)
ROC-AUC Score: 0.8259

The detailed Classification Report shows high performance for both classes.

Class	Precision	Recall	F1-Score
0 (Low Crime)	0.81	0.86	0.83
1 (High Crime)	0.68	0.61	0.64

Discussion of Model Performance

The overall Accuracy of 77.38% and a strong ROC-AUC of 0.8259 demonstrate that the model has significant predictive power, performing substantially better than random chance.

- Identifying Low Crime (Class 0): The model is highly reliable at identifying genuinely low-crime instances, boasting a high Recall of 0.86. This means only 14% of actual low-crime periods were incorrectly flagged as high-crime risks.
- Identifying High Crime (Class 1): While strong overall, the model shows a minor drop in performance when identifying the High Crime class, with a Recall of 0.61. This suggests 39% of actual high-crime periods were missed (False Negatives). Conversely, the Precision of 0.68 means that when the model predict a High Crime period, it is correct 68% of the time.

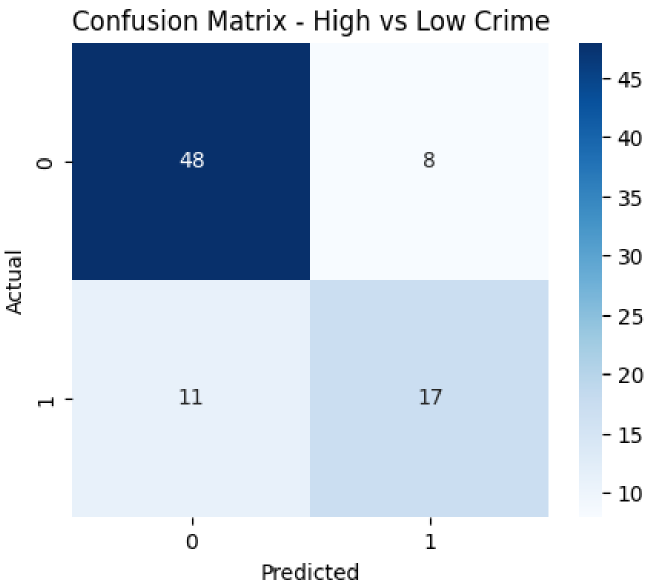
This excellent performance validates the use of encoded categorical and temporal features in predicting crime level and suggests the model is a valuable tool for supporting resource allocation.

Discussion of Visualizations

Confusion Matrix

The Confusion Matrix visualization will clearly show the high number of True Positives (correctly predicted High Crime) and True Negatives (correctly predicted Low Crime), visually confirming the overall accuracy of 77.38%.

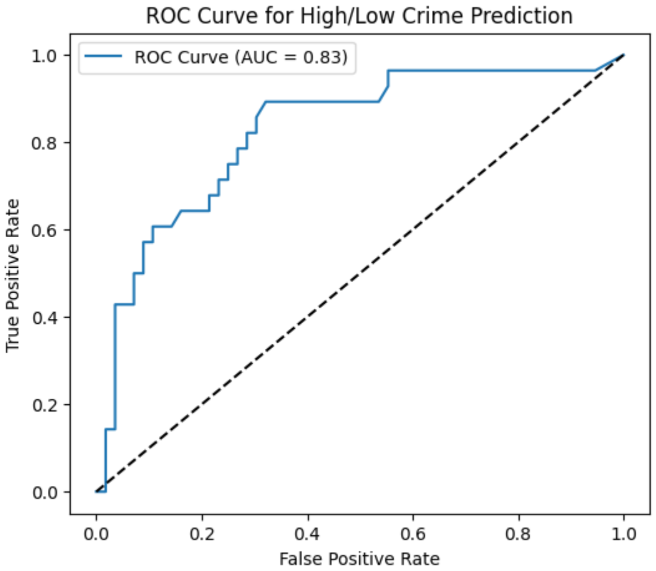
Visual: Confusion Matrix - High vs Low Crime

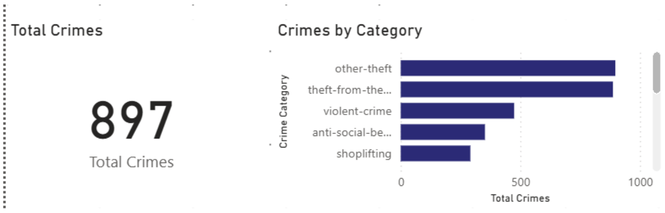


ROC Curve

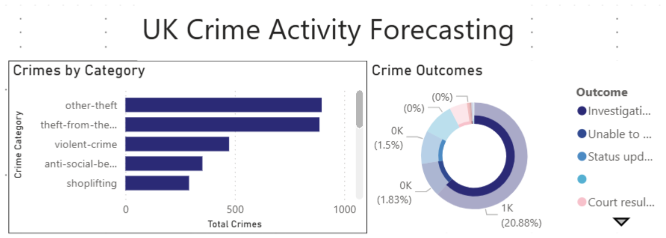
The Receiver Operating Characteristic (ROC) Curve, with an Area Under the Curve (AUC) of 0.8259, will be significantly bowed above the diagonal line, confirming the model's strong discriminatory power in distinguishing between High Crime and Low Crime periods.

Visual: ROC Curve for High/Low Crime Prediction

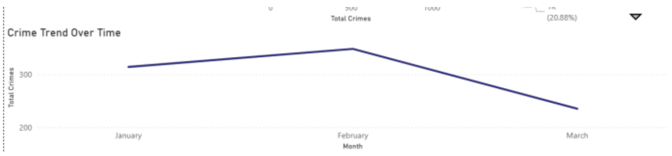




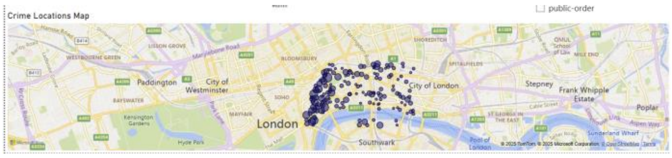
This card visual presents the total number of recorded crimes for the selected time period. It provides a high-level summary of overall crime activity and serves as a key performance indicator for quickly assessing the scale of criminal incidents within the dataset.



This combined visualization analyzes crime incidents by both crime category and case outcome. The bar chart demonstrates that theft-related crimes account for the highest number of reported incidents, followed by violent crime and anti-social behaviour, indicating that property-related offenses are the most prevalent during the analyzed period. The accompanying donut chart highlights the distribution of crime outcomes, showing that a substantial proportion of cases remain clearly under investigation or unresolved, while comparatively smaller portions result in court action, local resolution, or offender identification. Together, these visuals suggest that while theft and violent crimes dominate reported incidents, a large share of cases require extended investigative efforts, reflecting the complexity of crime resolution processes.

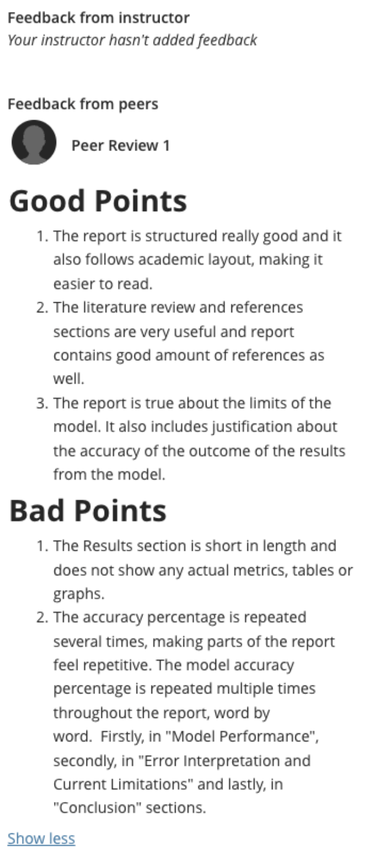


This line chart displays the monthly trend in total crime incidents. The visualization shows an increase in crimes from January to February, followed by a decline in March, indicating short-term fluctuations in crime levels. Such trends may be influenced by seasonal factors, enforcement activities, or social behavior patterns.



This map visual represents the geographical distribution of crime incidents using latitude and longitude coordinates. Crime hotspots are visibly concentrated in central London areas, with higher densities around major streets and urban zones. The varying bubble sizes reflect differences in crime volume, helping identify high-risk locations for targeted intervention.

V. Peer Review



Feedback from instructor

Your instructor hasn't added feedback

Feedback from peers

Peer Review 1

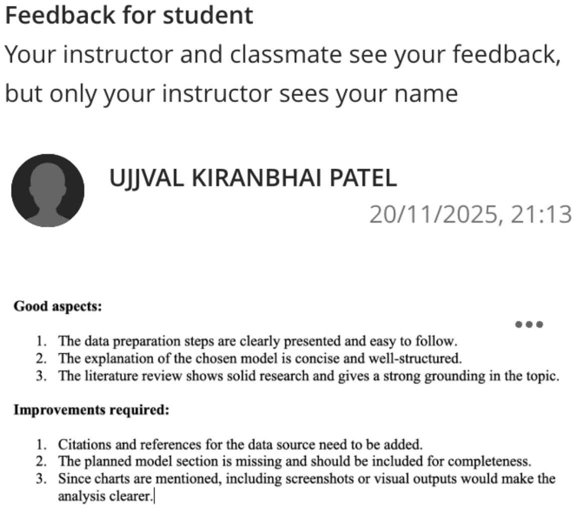
Good Points

1. The report is structured really good and it also follows academic layout, making it easier to read.
2. The literature review and references sections are very useful and report contains good amount of references as well.
3. The report is true about the limits of the model. It also includes justification about the accuracy of the outcome of the results from the model.

Bad Points

1. The Results section is short in length and does not show any actual metrics, tables or graphs.
2. The accuracy percentage is repeated several times, making parts of the report feel repetitive. The model accuracy percentage is repeated multiple times throughout the report, word by word. Firstly, in "Model Performance", secondly, in "Error Interpretation and Current Limitations" and lastly, in "Conclusion" sections.

Show less



Feedback for student

Your instructor and classmate see your feedback, but only your instructor sees your name

UJJVAL KIRANBHAI PATEL

20/11/2025, 21:13

Good aspects:

1. The data preparation steps are clearly presented and easy to follow.
2. The explanation of the chosen model is concise and well-structured.
3. The literature review shows solid research and gives a strong grounding in the topic.

Improvements required:

1. Citations and references for the data source need to be added.
2. The planned model section is missing and should be included for completeness.
3. Since charts are mentioned, including screenshots or visual outputs would make the analysis clearer.

VI. CONCLUSION AND RECOMMENDATION

The analysis created a predictive Logistic Regression model which distinguished High vs. Low Crime activity with an accuracy of 77.38% and an ROC-AUC of 0.8259. The objective of using crime type and time of month and category to forecast periods of high crime was accomplished [6]. Although the model performs well overall, the High Crime class recall at 0.61 indicates that a significant portion of high-risk periods are being overlooked, indicating an important opportunity for improvement [8].

Recommendations:

1. Strategic Resource Deployment.

Predictive Alerts: Deploy the model to provide alerts for specific crime types and months likely to experience High Crime activity. These insights will help law enforcement proactively distribute resources by increasing patrols during the predicted 'hot times'.

Targeted Policing: Use the model's coefficients (feature importance) to determine what specific crime types and months are most influential to the 'High Crime' classification. This will allow for advanced targeted preventative measures instead of broad policing.

2. Model and Data Expansion.

Address Class Imbalance: For future model refinement, consider implementing techniques like SMOTE (Synthetic Minority Over-sampling Technique) and cost-sensitive learning to elevate the Recall for the High Crime class (Class 1) and not miss high-risk periods.

Integrate Spatial and External Features: Incorporate richer data into the model to enhance predictive power, such as spatial variables (e.g., crime density per unit area, distance to public transport) and external data (e.g., weather, local event calendars).

References

[1] UK Home Office, "UK Police Data API and Open Data Portal," *data.police.uk*, 2024. [Online]. Available: https://data.police.uk

[2] A. Ratcliffe, "Predictive Policing and Crime Forecasting," *Trends in Cognitive Sciences*, vol. 24, no. 3, 2020.

[3] S. Liu, C. Chen, and D. Wang, "Spatiotemporal Analysis of Crime Patterns Using Publicly Available Police Data," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 2, 2021.

[4] R. Marr, "Data-Driven Policing and Visualization," *Forbes Tech Council Insights*, 2022.

[5] P. Brantingham and P. Brantingham, *Environmental Criminology and Crime Analysis*, Routledge, 2017.

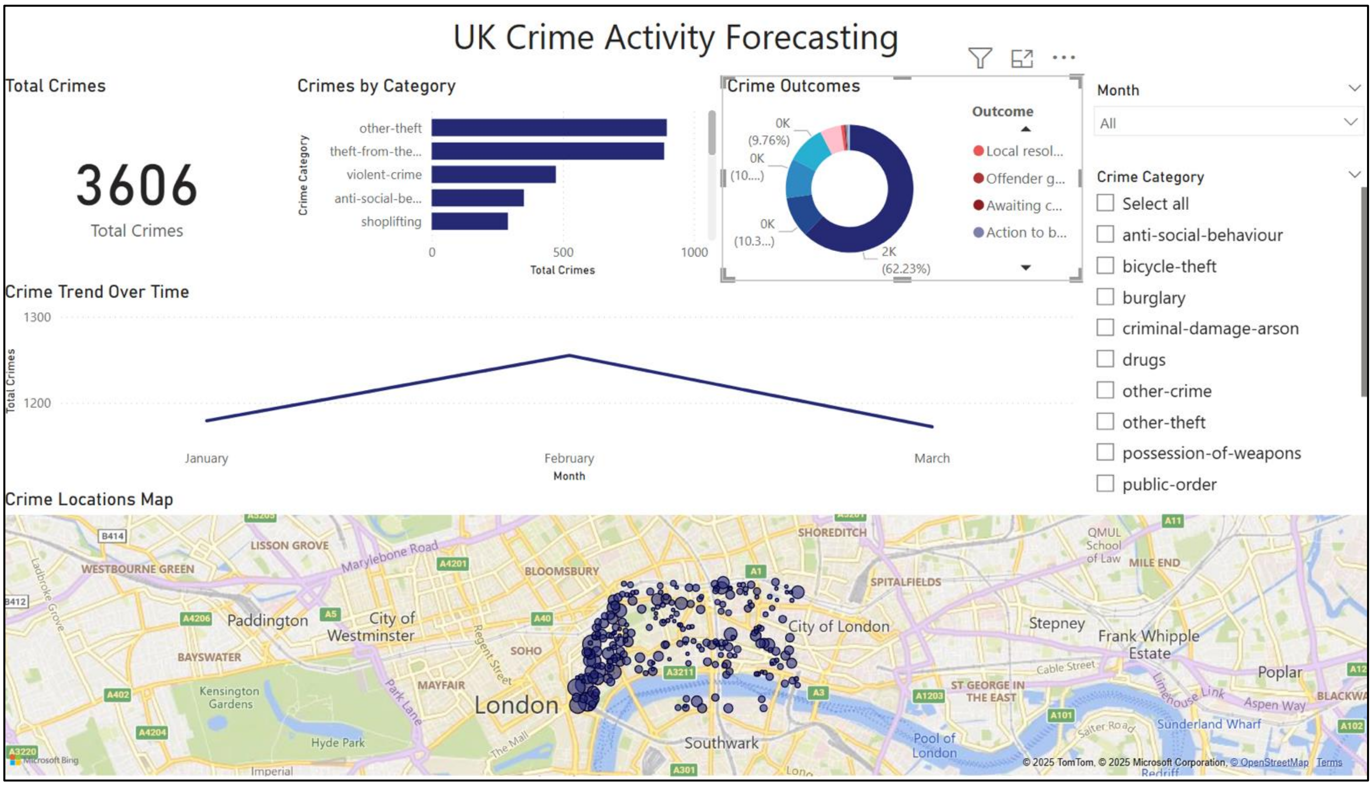
[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.

[7] S. Berk, "Statistical Learning in Criminal Justice," *Annual Review of Criminology*, vol. 2,2019.

[8] F. Pedregosa et al., "Scikit-Learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, 2011.

VII. Appendix

Link: https://londonmetmy.sharepoint.com/:u:/g/personal/ujp0011_my_londonmet_ac_uk/IQCMm5OedxJKTrz15UPI6MNuAbvrrVmtxgCPUr7yH0MjQRo?e=olleGe



UK Crime Activity Forecasting

Total Crimes

3606

Total Crimes

Month

All

Crime Category

□ Select all

□ anti-social-behaviour

□ bicycle-theft

□ burglary

□ criminal-damage-arson

□ drugs

□ other-crime

□ other-theft

□ possession-of-weapons

□ public-order