Prepared by:

Marina Mnoyan

Transforming Neighborhoods by Predicting Crime Outcomes in the UK

04/07/2022

Hyperlinked Content Table

1. Research Question

What crime-related and demographic factors influence whether crimes are solved in the UK?
Note: For the purpose of this question, we assume that crimes where the suspect is not identified have a lower likelihood of being solved.

Our initial hypothesis is that crime type-related features will have the highest impact on whether the suspect is identified or not (e.g., bike thefts generally have a lower incidence of suspect being identified compared to crimes related to public order disturbance or weapon possession). We also hypothesise that there may be disparities based on geography and neighborhood-level demographic factors.

2. Societal Value

Recently, there has been a lot of research and open conversation around systemic inequities and injustices in the democratic societies, many linked to demographic factors and social constructs such as race and age. These inequities impede the growth of our societies and limit the ability of each member of the society to reach their potential and live a dignified life.

Our research attempts to identify the factors that influence crime outcomes, with the goal of uncovering insights and creating recommendations for the Government of UK to address any potential inequities in this area.

3. Data Sources

Multiple data sources were used for this project. Key sources include the UK Police, Office of National Statistics, Open Data Portal and UK Data Service.

a. Crime level data (main dataset)

- Total of 1,513 CSV files spanning between April 2019 to March 2022
- Data is updated on a monthly basis so there is an opportunity to gather more data from the same source to validate the model and answer different research questions
- Data quality issues included missing values and size of the file which caused multiple system crashes

b. LSOA (Lower Layer Super Output Area) neighborhood level supplementary datasets

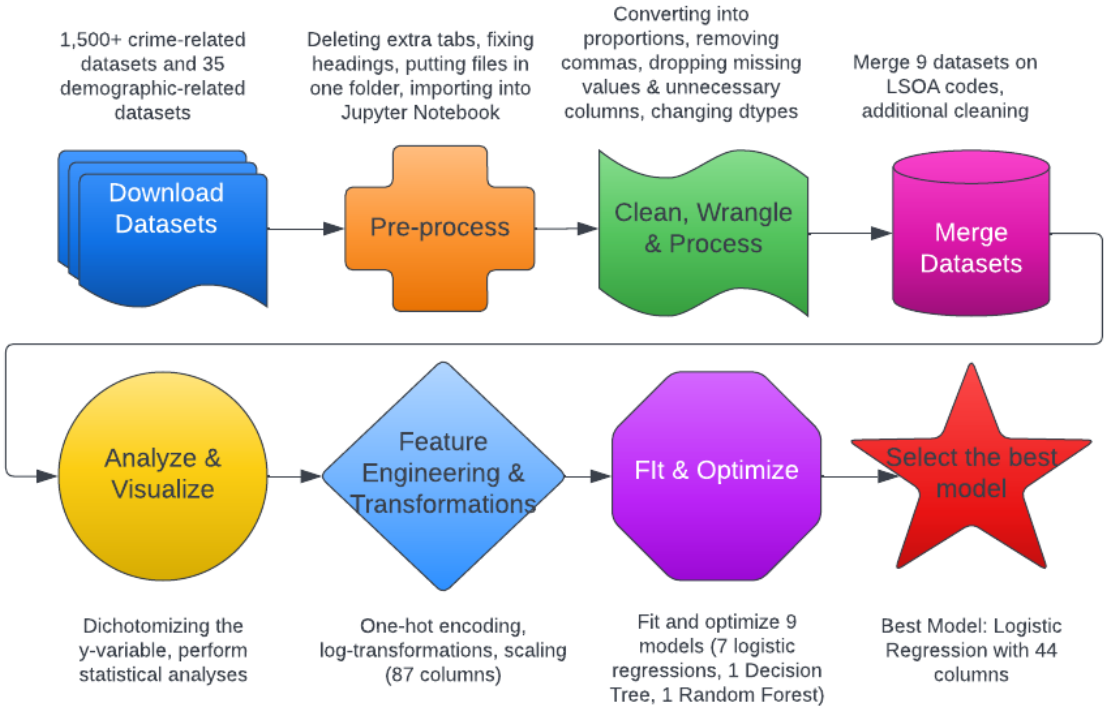
- This data was used to supplement crime data with demographic information on neighborhood (LSOA) level
- Datasets included information from 2011 Census; we make an assumption that the data is still valid given that newer Census data is not yet available in public domain
- A total of 35 datasets were downloaded but only 11 of them were deemed to be useful for this project

4. Project Framework & Deliverables

STEP	DELIVERABLE & TOOLS
1. Data acquisition & pre-processing	Google Chrome, MS Excel, Git Bash
2. Main dataset wrangling, processing & cleaning	Jupyter Notebook 1
3. Additional dataset Merging, Wrangling, Processing & Cleaning	Jupyter Notebook 2
4. Exploratory Data Analysis & Visualizations	Jupyter Notebook 3, Tableau Book
5. Feature Engineering	Jupyter Notebook 4
6. Modeling & Evaluation	Jupyter Notebook 5
7. Project Overview & Recommendations	PDF Report ←*current document* PDF MS PowerPoint
8. Technical Information	ReadMe text file

5. Process Summary

The following process was undertaken to complete the project:



6. Summary of Models

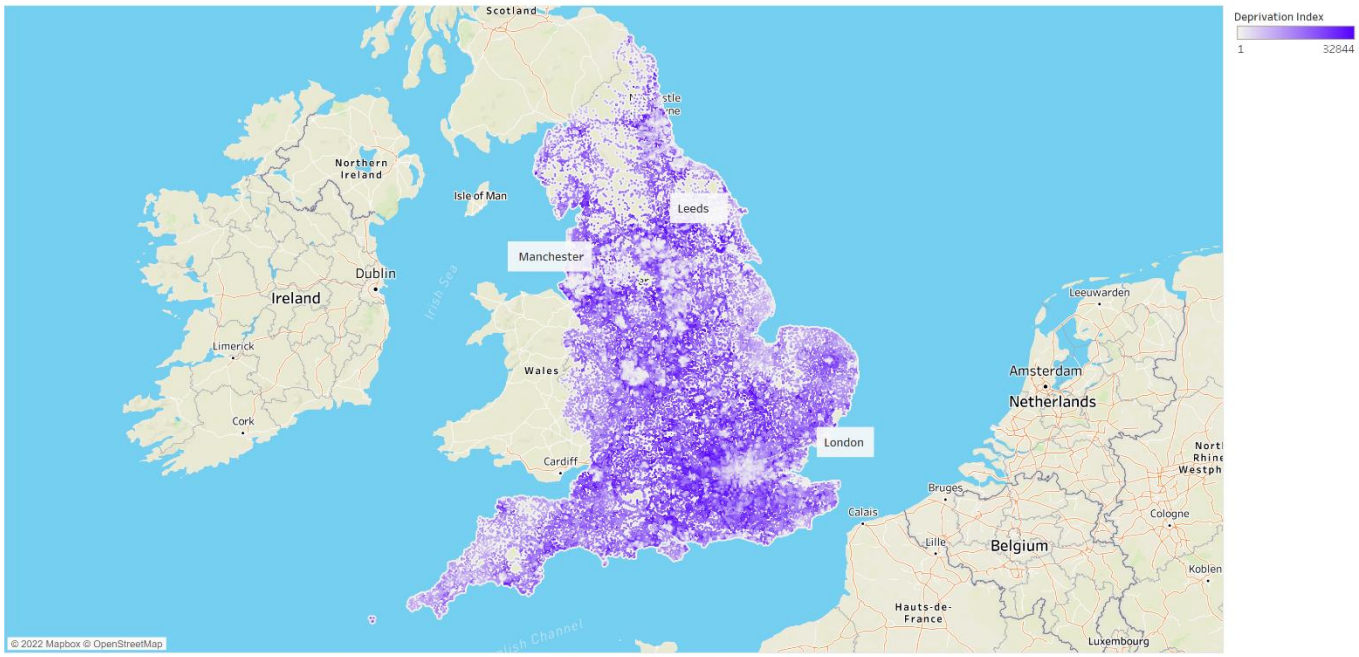
	Name	Transformation	Optimization	Features	Accuracy on Validation	F1-score on Validation
0	Base Logistic Model	None	None	87	61.45	69.17
1	Logistic Model 1	Scaled data	None	87	81.32	83.06
2	Logistic Model 2	Scaled data	C=0.1	87	81.32	83.05
3	Logistic Model 3	Scaled & log-transformed data	C=0.1	87	81.32	83.06
4	Logistic Model 4	Scaled data	C=0.1	81 (VT)	81.33	83.06
5	Logistic Model 5	Scaled data	C=0.1	44 (VT & Feature Selection)	81.38	83.10
6	Logistic Model 6	Scaled data	C=0.1	28 (KBest)	63.58	71.56
7	Decision Tree	Scaled data (optional)	max_depth = 8	87	81.17	82.55
8	Random Forest	None	n_estimators=31, max_depth=5	87	78.47	81.29

A total of 9 models were designed and evaluated. Given our research question, model simplicity, interpretability and high accuracy were our main model performance criteria. Based on these criteria, the Logistic Model 5 which used MinMax scaled data, optimization parameter C=0.1 and 44 features selected through variance thresholding and p-value/coefficient selection performed the best with **82%** accuracy and **83%** F1 score.

7. Findings from Exploratory Data Analysis

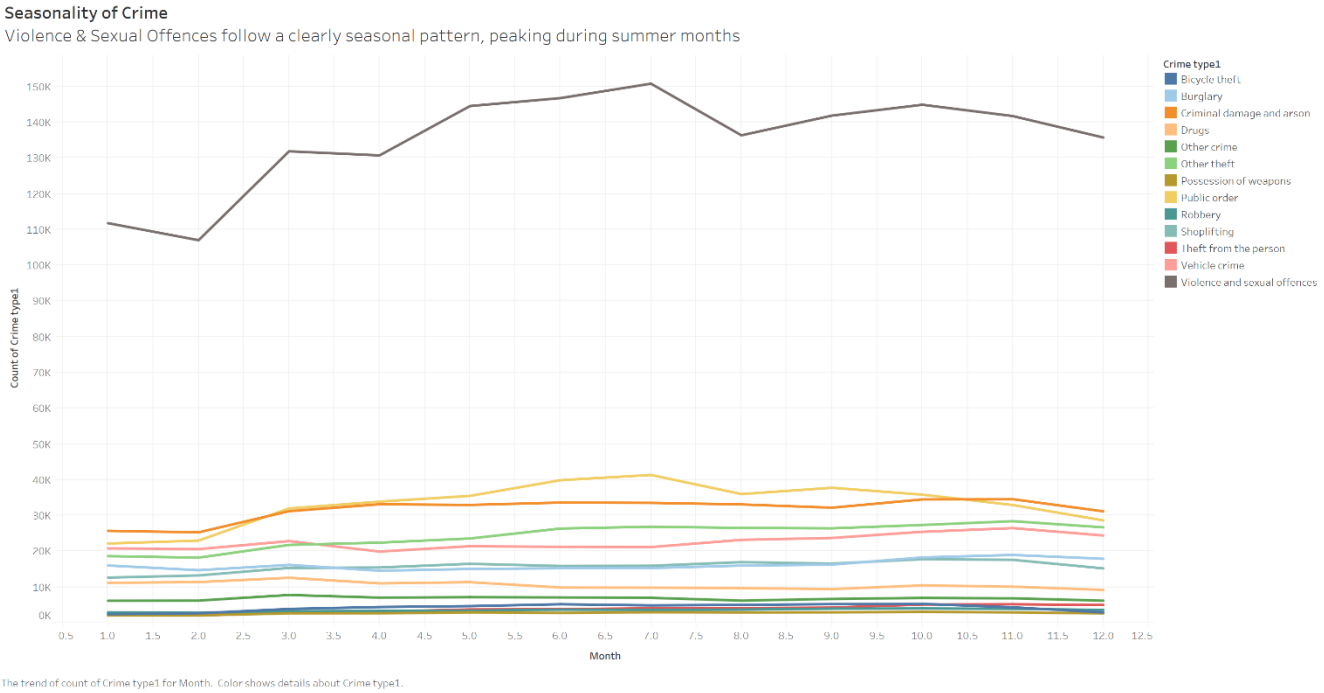
a) Poverty in the UK is concentrated in urban areas:

Poverty map across England and Wales (as measured by Deprivation Index where 1=Most Deprivated)
Poor areas are spread across the country, mostly concentrated around large cities

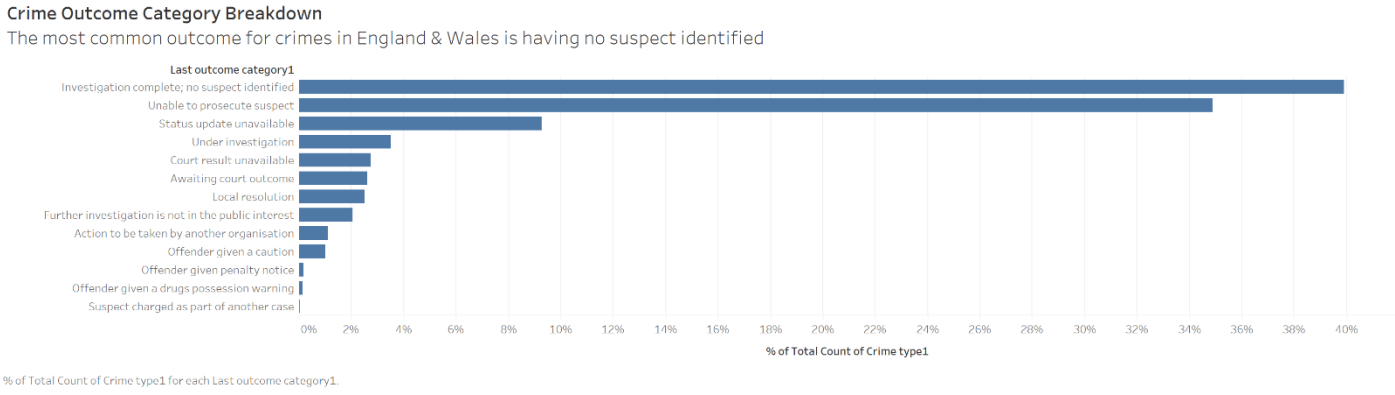


Map based on Longitude1 and Latitude1. Color shows details about Deprivation Index. The view is filtered on Exclusions (Latitude,Longitude), which keeps 472,313 members.

b) Violence & Sexual Offences Crimes follow a seasonal pattern, unlike most other crime types:

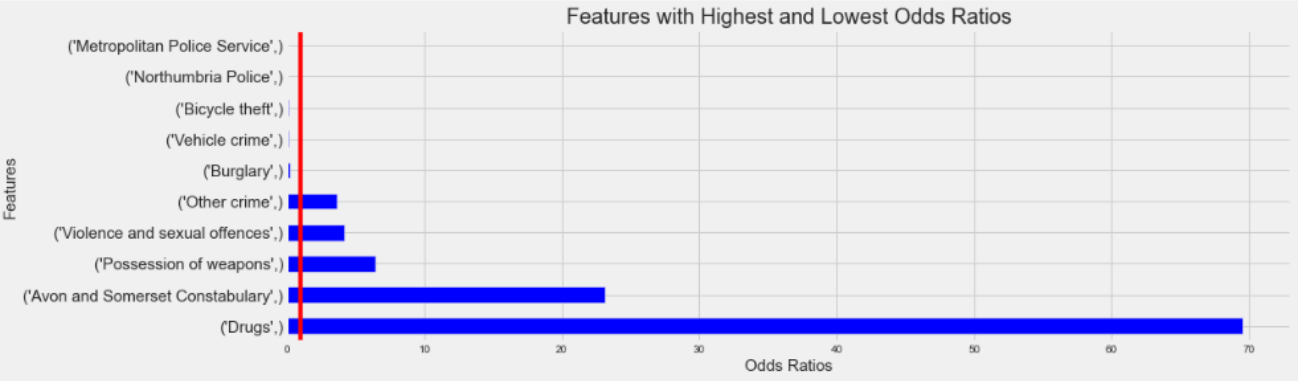


c) The most common crime outcome in the UK is suspect not being identified



8. Findings from Modeling

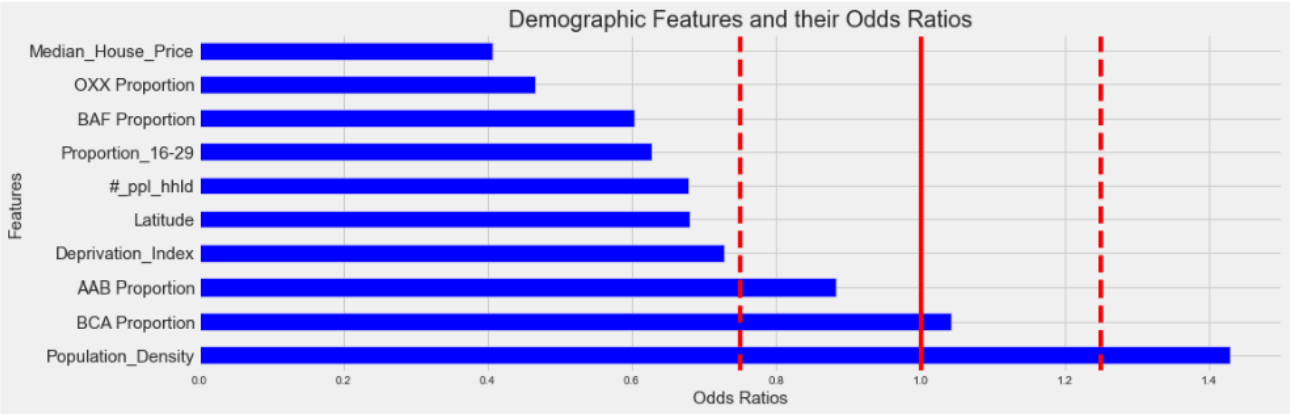
a) Factors that have the highest impact on whether the suspect will be identified or not are related to Crime types and disparities in various police jurisdictions



As hypothesized in the beginning of the project, some of the most significant predictors of whether a suspect will be identified or not are crime types. Unsurprisingly, crimes where suspects are easily identified are those related to possession offences, while the ones that have a harder time identifying suspects are the ones related to theft.

It is notable to see significant geographic differences, with different police jurisdictions having a big impact on the odds ratio on the outcome whether a suspect being identified or not. It is recommended that the UK Government focus on extra training and support of police forces in affected regions, more specifically Northumbria Police and Metropolitan Police Service given their poor track record in solving crimes.

b) Demographic factors influence whether the suspect will be identified or not



By looking at the data, we can see that:

There is a **lower** likelihood of suspect being identified in neighborhoods where…

- Median House Prices and Deprivation Indexes are high, possibly indicating underpolicing in affluent neighborhoods, or that the suspects are not from those neighborhoods
- There is a high proportion of Black African and Other Ethnicities, possibly indicating lower level of police involvement in these neighborhoods
- Residents are younger (higher proportion of 16-29 year olds), possibly indicating that these individuals may not have prior police engagement and therefore not being as easy to identify
- There is a higher number of people in household, possibly indicating poorer neighborhoods with multi-generational settings
- Latitude is higher, possibly indicating geographic disparities in the North

There is a **higher** likelihood of suspect being identified in neighborhoods where…

- Population density is high, possibly indicating a more trained police force in more populated urban areas

9. Summary and Potential Next Steps

In summary, we recommend that the UK Government look at the findings from this report in order to focused training to police service forces with a poor track record of solving crimes. They are also urged to commission further studies to identify the worrying signs of overpolicing and underpolicing inequities across the country, as highlighted by the impact of neighborhood-level demographic factors.

In the future, we could add search & order (also known as carding) data to this dataset to confirm our hypothesis around overpolicing in neighborhoods with racialized residents and lower socio-economic status.