Graffiti Prevalence Analysis in Urban and Rural Areas

A Data Analytics Project

Name: Hikmat Fadesewa Adedeji
Tools: R, Statistical Modelling
Dataset: Crime Survey for England and Wales (2013)
Outcome Variable: Prevalence of graffiti in residential areas

Executive Summary

This analysis evaluates the prevalence of graffiti in metropolitan versus rural areas, identifies other factors contributing to graffiti occurrences, then predicts out-of-sample. Utilising the 2013 crime survey data, it aims to inform strategies for resource allocation to enhance quality of life.

The analysis revealed that 30% of the area has prevalent graffiti, simple and multiple probability models explored the relationship between other factors and the likelihood of graffiti in the areas. The result showed that graffiti is common in urban areas and inclusion of confounding factors improved the model's fit.

The out-of-sample prediction was done using two models, logistic regression Predicts that 37.5% of areas will likely have graffiti prevalence, with urban, littered, and poor housing conditions as significant factors. The model's accuracy is 92.4%. Random Forest Performed similarly to the logistic model, identifying key variables like poor housing, littered areas, and deprivation as critical predictors. With an accuracy of 92.1%. The two models were compared and the preferred model was the logistic model due to the model's ability to provide clearer insights into the significance of variables which is valuable for making informed decisions. However, the model doesn't capture complex interactions, the analysis proposed using model refinement and including relevant data that could act as predictors to improve the model.

Considering the findings from the analysis the following recommendations were made: resource allocation, community engagement, enhancing housing conditions, organising cleaning campaigns, and enhancing public spaces. Employing these data-driven insights will enhance the quality of life in affected areas.

Introduction

Examining the prevalence of graffiti is crucial for understanding its impact on neighbourhood appearance, safety perception and socio-economic conditions. This analysis aims to assess the prevalence of graffiti in metropolitan areas relative to the countryside. Additionally, it considered other factors that could contribute to the prevalence of graffiti and predicted out-of-sample. It uses the 2013 crime survey data to provide valuable insights that can be used to develop effective strategies for resource allocation to improve the quality of life in affected areas.

Data preparation

The crime survey data has 8843 observations and 32 variables. Relevant variables for the analysis were selected, and new variables were created to facilitate a seamless analysis and to incorporate factors that would enhance the quality of the analysis. A detailed explanation of this process is provided in Appendix 1a. The outcome variable, “vandcomm,” indicates the prevalence of graffiti in the area. The primary explanatory variable, “rural2” identifies whether the area is rural or urban, allowing for the comparison between metropolitan and countryside.

Appendix 1 summarises the other variables used in the analysis. After the cleaning and transformation, 8086 observations and 20 variables were selected representing 91% of the original dataset. this reduction in observations may impact data representativeness and generalizability.

Analysis

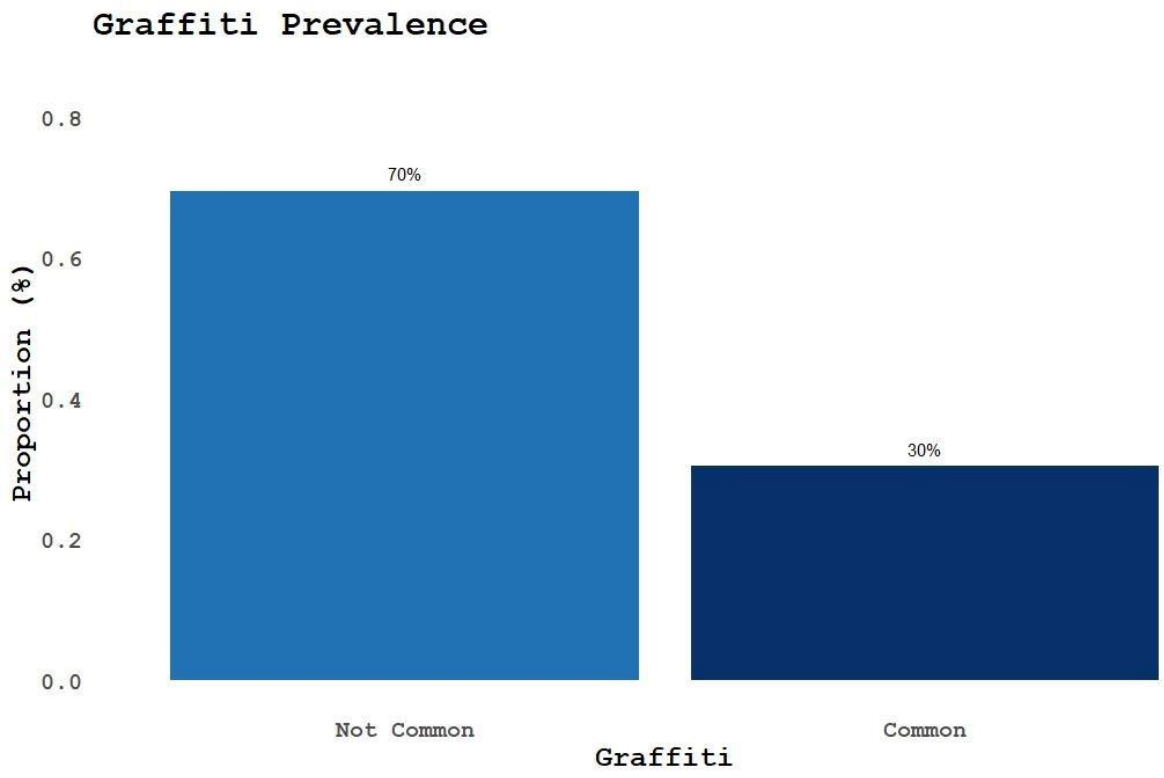

Figure 1: Graffiti Prevalence

Figure 1 depicts the prevalence of graffiti, the chart reveals that 30% signifies areas where graffiti is prevalent. The succeeding figure provides a comparative analysis of graffiti between rural and urban areas.

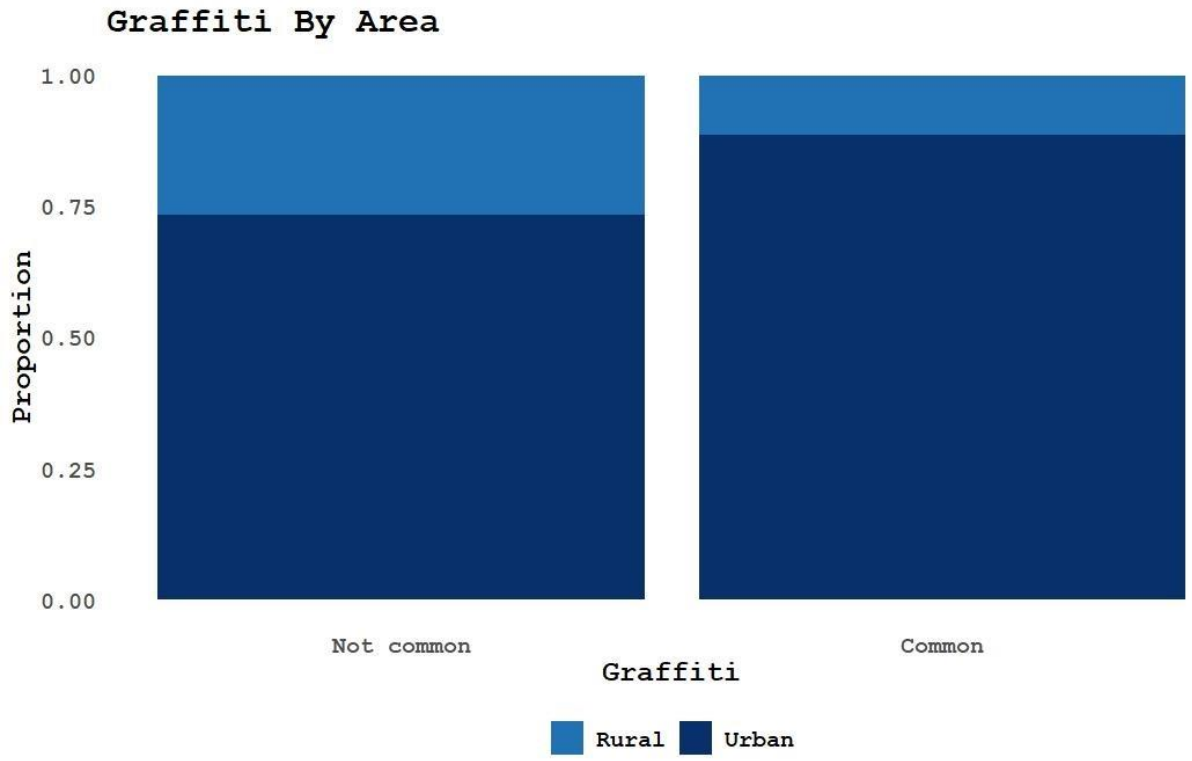

Figure 2:

Figure 2 illustrates that most areas where graffiti is prevalent are in the metropolitan area, which, in this context, is classified as urban. The subsequent analysis employs a linear probability model.

Table 1:

LINEAR PROBABILITY MODEL		
VARIABLES	**MODEL 1**	**MODEL 2**

	Coef.(p-value)	**Coef. (p-value)**
Urban	0.190*** (0.012)	0.022*** (0.008)
Littered		0.255*** (0.009)
Poor_house		0.523*** (0.009)
Owned		-0.013 (0.019)
Rented		0.003 (0.019)
Age_16_24		-0.006 (0.013)
Age_25_34		-0.006 (0.010)
Age_35_44		-0.016 (0.010)
Age_45_54		-0.002 (0.010)
Most_deprived		-0.030*** (0.007)
Ethnic_Mixed		-0.018 (0.030)
Ethnic_Black_Bri		0.038** (0.017)
Ethnic_Asian_Bri		0.041*** (0.015)
Ethnic_Chinese		0.068** (0.028)
Male		0.004 (0.006)
Crime_Victim		-0.014 (0.009)
working		-0.002 (0.007)
Constant	0.156***	(0.011) (0.021)
Observations	8,086	8,086
R2	0.029	0.639
Adjusted R2	0.029	0.638
Residual Std. Error	0.453 (df = 8084)	0.277 (df = 8068)
F Statistic	243.277*** (df = 1; 8084)	840.143*** (df = 17; 8068)

Table 1 presents both simple and multiple linear probability models. Model 1 reveals that living in an urban area increases the probability of graffiti prevalence by 19%,

with an R-squared of 2%. In contrast, Model 2 accounts for additional factors that can affect the prevalence of graffiti, achieving an R-squared of 63%, indicating a better fit. The subsequent analysis aims to predict out-of-sample performance.

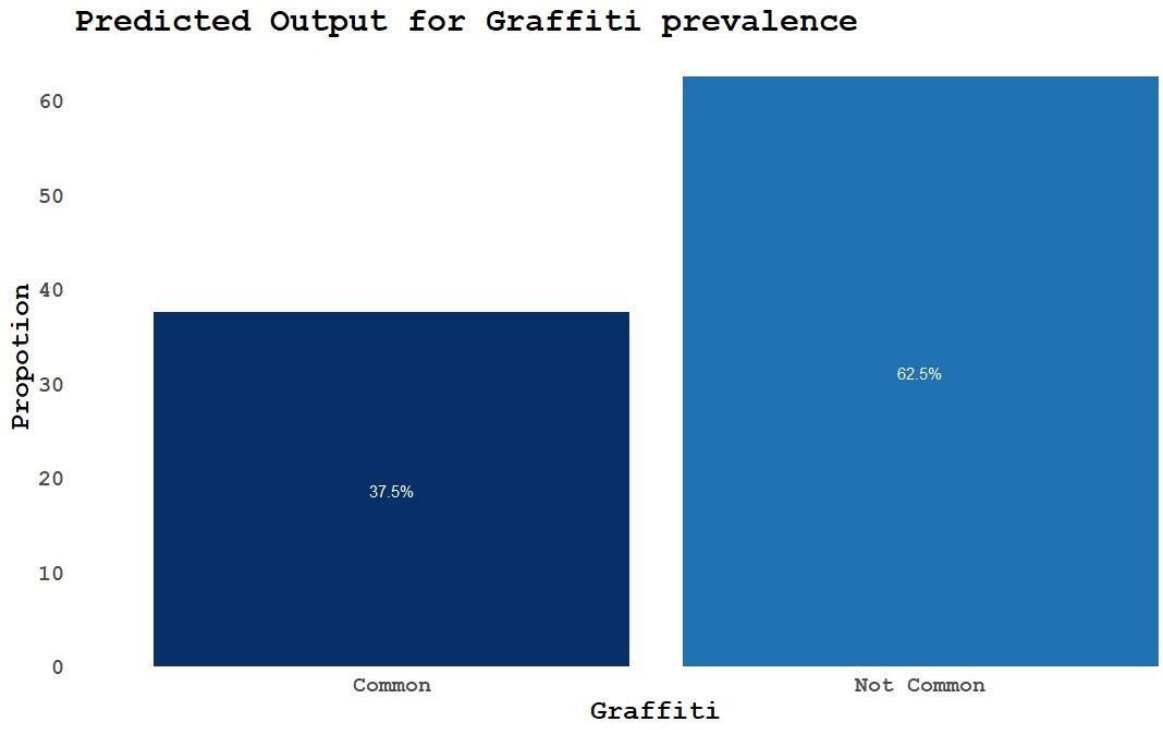

Figure 3

The prediction analysis employs logistic regression due to its suitability for analyzing binary outcome variables with binary predictors. The dataset was split into training and testing sets using an 80:20 ratio. The model, trained on this dataset, predicts that 37.5% of the areas are likely to have prevalent graffiti, as shown in Figure 3

Table 2

LOGIT REGRESSION OUTPUT	
	MODEL 1
VARIABLES	**AME(P-VALUE)**
Urban	0.029*** (0.010)
Littered	0.258*** (0.021)
Poor_house	0.249*** (0.027)
Owned	-0.022 (0.019)
Rented	-0.005 (0.019)
Age_16_24	-0.006 (0.013)
Age_25_34	-0.009 (0.010)
Age_35_44	-0.016 (0.010)
Age_45_54	-0.002 (0.010)
Most_deprived	-0.023*** (0.007)
Ethnic_Mixed	-0.015 (0.027)
Ethnic_Black_Bri	0.024 (0.016)
Ethnic_Asian_Bri	0.024* (0.014)
Ethnic_Chinese	0.080** (0.031)
Male	0.008 (0.006)
Crime_Victim	-0.018** (0.009)
working	-0.002 (0.007)
Constant	-0.409*** (0.040)
Observations	6,469

Akaike Inf. Crit.	2,839.086

In Table 2, the logistic regression model suggests that urban areas, littered environments, and poor housing conditions significantly increase the probability of graffiti prevalence. The model also indicates that certain racial demographics are associated with higher probabilities of graffiti prevalence, while the most deprived areas and being a crime victim are associated with lower probabilities.

Table 3

LOGIT CONFUSION MATRIX		
	Notcommon	common
Common	105	500
Notcommon	995	17

In table 3 above the model efficiently predicted 500 cases of areas where graffiti is common and 995 cases as not common areas, leading to an overall accuracy of 92.4%. This suggests that the model is expected to predict the outcome correctly approximately 92.4% of out-of-sample data.

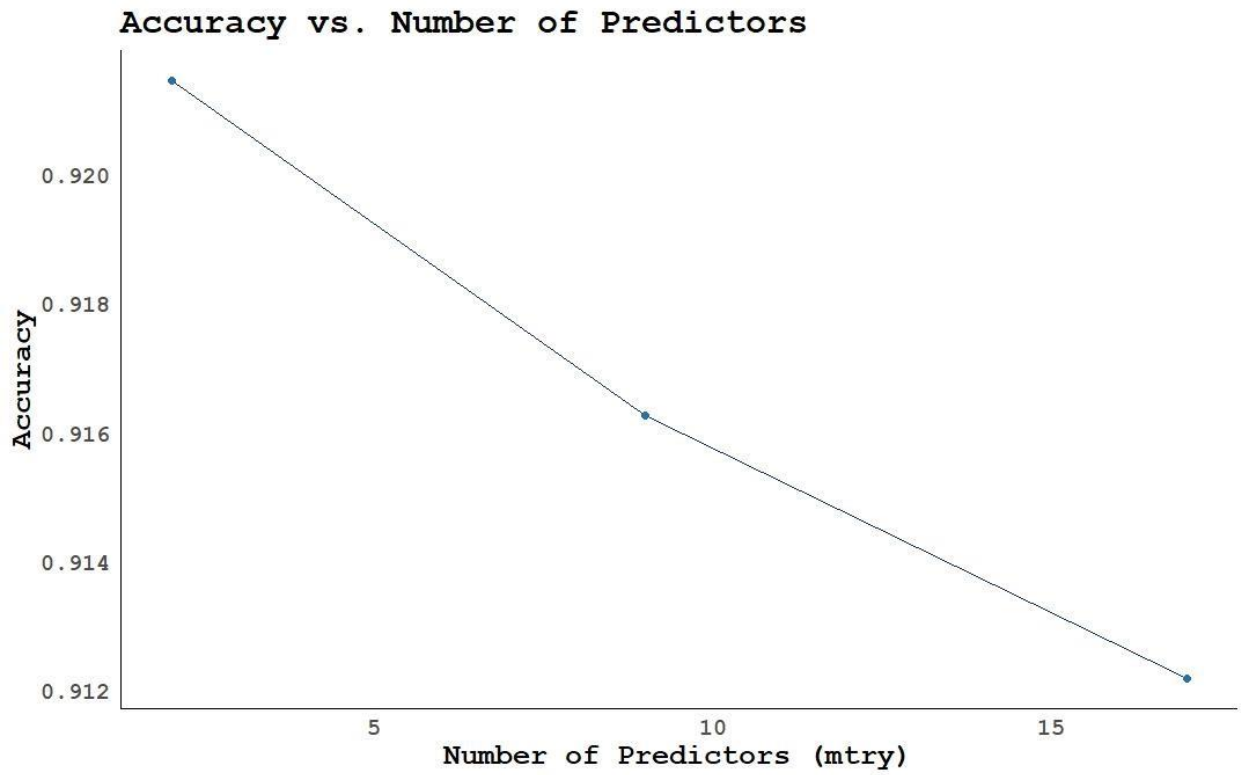

Figure 4: Accuracy VS. Prediction

The Random Forest model performs best with mtry 2, achieving the highest accuracy at 92.1%. Increasing mtry to 9 and 17 results in slight performance declines, with accuracy dropping to
91.6% and 91.2%, respectively.

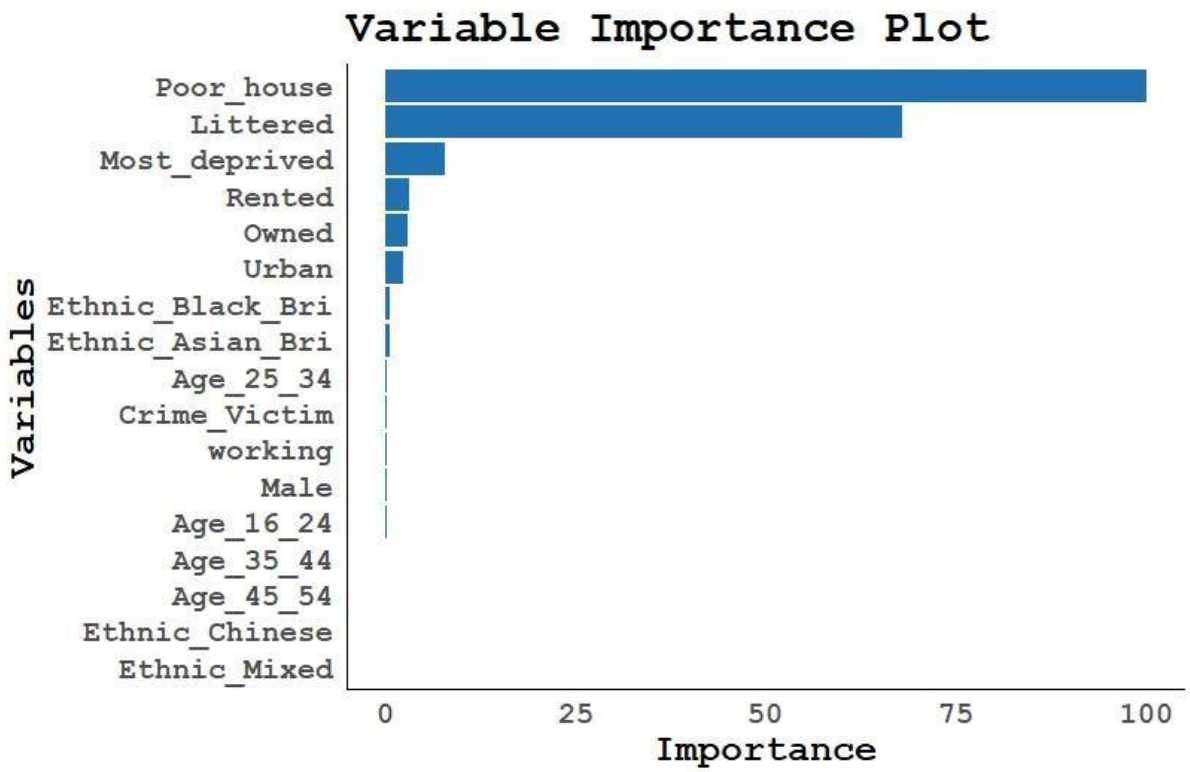

Figure 5:Variable Importance plot

The most important variables for predicting the prevalence of graffiti were selected manually and automatically selected, the manually selected variables are shown in Appendix 4 Variables like poorhouse, littered, and most deprived have low AICs, indicating their importance in predicting the outcome variable. The generated variables, as displayed in the above chart, also highlight the significance of these variables and those that are not statistically significant in predicting the outcome variable. Variables like poorhouse, littered, and most deprived have low AICs, indicating their importance in predicting the outcome variable.

Logit Vs. Random Forest

The logit and Random forest models produced similar accuracy results. The logit model achieved an accuracy of 92.4% which is slightly higher than that of random forest. Logit model has a marginally better performance in terms of the proportion of correct predictions.

The Logit model uses AME and p-values for significance, while Random Forest uses importance scores, handling complex interactions better.

The Logit model is simpler and more interpretable; the Random Forest is more complex and captures intricate patterns like multicollinearities.

Key variables for Random Forest are "Poor house," "Littered," and "Most deprived." For Logit, they are "Littered," "Poor_house," "Urban," "Most_deprived," and "Ethnic_Chinese."

Conclusion

The Logit model achieved 92.4% and provides clearer insights into the significance of each variable which is valuable for making informed decisions regarding areas where graffiti is common, this makes it a better model. However, it doesn’t capture complex interactions, that can be improved by using Ridge and Elastic Net regularization to handle multicollinearity and shrink less important coefficients. Adding more relevant data, expanding the sample size, incorporating interaction terms, polynomial features, and geospatial, and temporal data can enhance performance. Iteratively adjusting regularization parameters and exploring ensemble methods will further improve the model.

Furthermore, the variables "Poor house," "Littered areas," and "Deprivation" are the most significant predictors of graffiti occurrence. This suggests that these factors critically influence whether graffiti will be present in a given area, whether urban or rural. To address this, enhance housing conditions, organize cleanliness campaigns, address deprivation, allocate resources to high-risk areas, enhance public spaces, engage the community, and use data-driven strategies.

APPENDICES

APPENDIX 1a: Data Preparation

This section outlines the steps taken to clean and prepare data for analysis.

Population

The crime survey for 2013 has 8843 observations and 32 variables.

Sample

After cleaning, we have 8086 observations and 20 variables, representing 91% of the original dataset. However, this reduction in observations may impact data representativeness and generalizability. Thus, while the refined dataset provides improved insights, caution is needed when interpreting results due to potential limitations from data reduction.

Data Transformation:

Variables indicating different degrees of commonness and non-common situations were transformed into a binary format. All degrees of commonness were grouped as “Common,” and non-common situations were grouped as “Not Common.” This transformation was done to ensure a seamless analysis by categorizing it into two broad categories, making the analysis easier. Frost (2013) supports this rationale by discussing Frost discusses how simplifying data by grouping variables can make analysis more actionable and interpretable for managers, leading to better decision-making.

The deprivation index variable was grouped by categorizing the first three quintiles as "most deprived." This strategic decision aligns with UK policy practices and academic research, ensuring that resources and interventions are directed towards the populations with the highest levels of deprivation.

Variable Selection:

Variables were selected to investigate the business questions based on the following themes: socioeconomic factors, demographics, safety perception, and deprivation. Even though some variables may not be statistically significant, they have theoretical importance. This approach is supported by Smith (2018), who discuss the importance of including theoretically relevant variables in regression models to avoid omitted variable bias and ensure comprehensive model specification.

Appendix 1:VARIABLE DEFINITION

VARIABLE	**DEFINITION**
Graffiti	Binary Variable with 1 indicating areas where graffiti is common and 0 otherwise
Graffiti2	Categorical variable with values "Common" to indicate areas where graffiti is common, and "Notcommon" indicating otherwise
Poor_house	Binary variable with 1 indicating areas with poor houses and 0 otherwise
Littered	Binary variable with 1 indicating littered areas and 0 otherwise
Urban	Binary variable with 1 indicating Urban areas and 0 Rural areas
Male	Binary Variable with 1 indicating male gender and 0 for female
Crime_victim	Binary Variable with 1 indicating crime victims in the past 12 months and 0 otherwise
Working	Binary Variable with 1 indicating people that are working and 0 otherwise
Most_deprived	Binary Variable with 1 indicating deprived areas and 0 otherwise
Owned	Binary Variable with 1 indicating those that own property and 0 otherwise
Rented	Binary variable with 1 indicating people that rented property and 0 otherwise
Age_16_24	Binary variable with 1 indicating those whose age is between 16 to 24 years and 0 otherwise
Age_25_34	Binary variable with 1 indicating those whose age is between 25 to 34 years and 0 otherwise

Age_35_44	Binary variable with 1 indicating those whose age is between 35 to 44 years and 0 otherwise
Age_45_54	Binary variable with 1 indicating those whose age is between 45 to 54 years and 0 otherwise
Ethnic_Mixed	Binary variable indicating those whose ethnic is mixed and 0 otherwise
Ethnic_Asian_Bri	Binary variable indicating those whose ethnic is Asian or British Asian and 0 otherwise
Ethnic_Black_Bri	Binary variable indicating those whose ethnic is Black or Black British and 0 otherwise
Ethnic_Chinese	Binary variable indicating those whose ethnic is Chinese and 0 otherwise

N = 8086

Appendix 2: DESCRIPTIVE STATISTICS

VARIABLE	**MEAN**	**STANDARD DEVIATION**	**MIN**	**MAX**
Graffiti	0.304	0.460	0	1
Male	0.457	0.498	0	1
Working	0.533	0.499	0	1
Most_deprived	0.617	0.486	0	1
Poor_house	0.395	0.489	0	1
Littered	0.467	0.499	0	1
Urban	0.780	0.414	0	1
Crime_Victim	0.157	0.364	0	1
Owned	0.627	0.484	0	1
Rented	0.341	0.474	0	1

Age_16_24	0.078	0.369	0	1
Age_25_34	0.155	0.362	0	1
Age_35_44	0.164	0.371	0	1
Age_45_54	0.172	0.378	0	1
Ethnic_Mixed	0.106	0.103	0	1
Ethnic_Asian_Bri	0.480	0.214	0	1
Ethnic_Black_Bri	0.035	0.184	0	1
Ethnic_Chinese	0.012	0.109	0	1

N = 8086

Appendix 3: Random Forest

RF MODEL PERFORMANCE METRIC		
mtry	**Accuracy**	**Kappa**
2	0.9214678	0.8233694
9	0.9162736	0.8108888
17	0.9121919	0.8008038

Appendix 4: AIC TABLE

NO	VARIABLE NAME	AIC
1	Urban	9682
2	Poor_house	4527
3	Littered	5250
4	Male	9940
5	Owned	9579
6	Rented	9587
7	Age_16_24	9916
8	Age_25_34	9897
9	Age_45_54	9942
10	Most_deprived	9204
11	Ethnic_Chinese	9933
12	Ethnic_Black_Bri	9855
13	Ethnic_Asian_Bri	9876
14	Ethnic_Mixed	9939
15	Crime_Victim	9935
16	working	9942

REFERENCES

Frost, J., 2013. The benefits of simplifying data: a management perspective. *Journal of Business Research Method,* 12(4), 233-245.

Smith, G., 2018. Step away from stepwise. *Journal of Big Data*. 5(32), 10.1186/s40537-018-0143-
6.