Identifying High-risk Areas for Theft in London using Machine Learning

Introduction

This study aims to find a machine learning model which perform better in predicting theft risk across London, while identifying key socioeconomic factors that influence high-risk area distributions.

Literature review

1) Yue, H. and Chen, J. (2025) 'Interpretable spatial machine learning for understanding spatial heterogeneity in factors affecting street theft crime', Applied Geography, 175, p. 103503. doi: 10.1016/j.apgeog.2024.103503.

This paper uses XGBoost machine learning algorithm combined with SHAP interpretation model to predict crime rate, and the results show that the proportion of non-local population and age group contribute the most to crime prediction.

2) González-Prieto, Á. et al. (2023) 'Hybrid machine learning methods for risk assessment in gender-based crime', Knowledge-Based Systems, 260, p. 110130. doi: 10.1016/j.knosys.2022.110130.

This research propose a mixed machine learning model that integrates Nearest Centroid and statistical methods to predict the risk of recidivism in gender-based violence cases. The results show that the hybrid model can improve the effectiveness of police protection by up to 25% comparing with existing risk assessment methods.

Research question

1. Compared to traditional statistical methods, do machine learning models significantly improve the accuracy of predicting the risk of theft across different areas of London?
2. Which socioeconomic factors have the greatest influence on the spatial distribution of high-risk areas for theft in London?

Data

1. LSOA London Crime Data 2010-2024
2. London Census Data: demography, housing, labor market etc.

Method

Q1

Machine Learning Approach: using XGBoost model to capture complex nonlinear relationships in the data
Statistical Approach: using Logistic regression as a baseline statistical model for comparison purposes.
Comparing accuracy、 F1-score、 AUC of both approaches, along with a paired t-test to assess whether the performance difference between models was statistically significant.

Q2

Feature Importance Assessment: using SHAP from XGBoost model to interpret the contribution of each variable to the prediction outcome
Spatial Analysis: using Getis-Ord Gi to identify the high-risk ares, to help better explain the results of SHAP

Results

Metrics for Model Evaluation:

Accuracy, F1-score, AUC

Visualization:

SHAP Summary Plot: Displays the overall importance of socioeconomic factors in predicting theft crime risk across different areas.

SHAP Dependence Plot: Explores the relationship between a key variable and the model output, revealing potential nonlinear effects.

Spatial Heat map (if applicable): Visualises the geographic distribution of SHAP values, indicating regions where specific factors have a stronger influence on theft crime rates.

Interpretation of the Results

Model Performance: XGBoost model or basic model

Key Influencing Factors: SHAP values identify the most influential socioeconomic variables driving theft crime risk

Spatial Variability: mapping SHAP values to uncover spatial heterogeneity in risk factors,