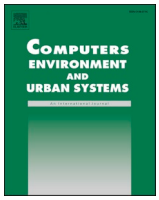
Computers, Environment and Urban Systems 94 (2022) 101789

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus



Interpretable machine learning models for crime prediction

Xu Zhang [a,b], Lin Liu [b,c,*], Minxuan Lan [d], Guangwen Song [b], Luzi Xiao [b], Jianguo Chen [b]

[a] School of Computer Sciences and Cyber Engineering, Guangzhou University, Guangzhou, China
[b] Center of Geoinformatics for Public Security, School of Geographic Sciences, Guangzhou University, Guangzhou, China
[c] Department of Geography, University of Cincinnati, Cincinnati, USA
[d] Department of Justice Sciences, The University of Findlay, Findlay, USA

ARTICLE INFO

Keywords:
Crime prediction
Machine learning
XGBoost
Model interpretability
SHAP value

ABSTRACT

The relationship between crime patterns and associated variables has drawn a lot of attention. These variables play a critical role in crime prediction. While traditional regression models are capable of revealing the contribution of the variables, they are not optimal for crime prediction. In contrast, machine learning models are more effective for crime prediction, but most of them cannot estimate the contribution of each individual variable. This study aims to overcome this limitation by taking advantage of the interpretability of advanced machine learning models. Based on the routine activity theory and crime pattern theory, this study selects 17 variables for the crime prediction. The XGBoost algorithm is adopted to train the prediction model. A post-hoc interpretable method, Shapley additive explanation (SHAP), is used to discern the contribution of individual variables. A variable with a higher SHAP value has a higher contribution to the crime prediction model. In addition to the global model for the entire area, a local model is calibrated at each study unit, revealing the spatial variation of the variables' unique contributions. Among all 17 variables used in this model, the proportion of the non-local population and the ambient population aged 25–44 contribute more than other variables in predicting crime. The more the ambient population aged 25–44 in the area, the more the public thefts. Additionally, local SHAP values are mapped to demonstrate each variable's contribution to the crime prediction model across the study area. The results of the local models can help the police tackle the most important factors at each location, while the global model identifies the important factors for the entire region.

1. Introduction

Machine learning technology has achieved great success in many fields and has been widely applied to some important practical tasks, such as face recognition (Taigman, Yang, Ranzato, & Wolf, 2014; Sun, Wang, & Tang, 2014), speech recognition (Deng, Hinton, & Kingsbury, 2013), automatic driving (Hoermann, Bach, & Dietmayer, 2018), intelligent medical analysis (Choy et al., 2018), etc. While machine learning outperforms other models in many tasks, this method generally lacks transparency and interpretability. Thus machine learning is typically regarded as a "black box" approach, and it is difficult to explain what happened in the "black box" (Apicella, Isgrò, Prevete, & Tamburrini, 2020).

Interpretability of machine learning has drawn attention from both social sciences and artificial intelligence fields (Miller, 2019). However, the interpretability and transparency of machine learning based crime prediction models remain questionable (Alves, Ribeiro, & Rodrigues, 2018; Rummens, Hardyns, & Pauwels, 2017; Wang, Ge, Li, & Chang, 2020; Zhang, Liu, Xiao, & Ji, 2020). Because of this, practitioners tend to distrust the results of such models. Therefore, there is an urgent need for an interpretable and transparent crime prediction model, so that practitioners would know the calculation logic and factors they need to pay attention to. Also, spatial variations of each variable in the machine learning model shall receive special attention as they influence crime opportunities (Lan, Liu, & Eck, 2021; Wilcox & Eck, 2011).

Before machine learning is introduced to crime prediction, traditional crime prediction models solely use historical crime data with the assumption that crime events are near-repeated in space and time. That is to say, spatiotemporal information of historical crime events is used to predict the distribution of criminal activities in a later time (Farrell & Pease, 1993; Sherman, Gartin, & Buerger, 1989). Some examples are: near-repeat prediction model (Townsley, 2003) and density estimation model (Chainey & Ratcliffe, 2013; Kalinic & Krisp, 2018). Such models are suitable for crime risk prediction in large spatial and temporal scales

* Corresponding author at: Center of Geoinformatics for Public Security, School of Geographic Sciences, Guangzhou University, Guangzhou, China.
E-mail addresses: lin.liu@uc.edu (L. Liu), minxuan.lan@findlay.edu (M. Lan), xiaoluzi@gzhu.edu.cn (L. Xiao), chenjg@gzhu.edu.cn (J. Chen).

https://doi.org/10.1016/j.compenvurbsys.2022.101789
Received 29 November 2021; Received in revised form 27 January 2022; Accepted 6 March 2022
Available online 12 March 2022

when environmental details and subject behavior information are insufficient. However, their prediction accuracy cannot be guaranteed at fine spatial and temporal scales.

As various environmental factors are found to influence crime opportunities, fine-scale crime predictions incorporating such factors become possible. Since the middle of the 20th century, with the cross integration and gradual development of criminology and geography, basic theories of criminal geography have emerged. Some examples are crime opportunity theory (Cohen, 1981), routine activity theory (Cohen & Felson, 1979), rational choice theory (Cornish & Clarke, 1987), and crime pattern theory (Brantingham & Brantingham, 1993a). Crime opportunity theory emphasizes that the distribution of people's daily activities and crime opportunities have an important impact on the spatial pattern of crime (Cohen, 1981; Weisburd, Lawton, & Ready, 2012). Crime opportunity is the key to influencing the offenders' choice to commit crimes or not (Cornish & Clarke, 1987). Routine activity theory argues that the convergence of motivated offenders, suitable targets/victims, and the absence of capable guardians at time and space is needed for a crime event to happen. People's routines may provide such convergences (Cohen & Felson, 1979). The rational choice theory also suggests the potential influence of built environment factors on offenders' rational thinking process (Stummvoll, 2009). Crime pattern theory focuses on places and opportunities, and especially emphasizes the overlap of consciousness space of offenders and victims. Offenders always tend to choose the places they are familiar with when searching for potential targets (Brantingham & Brantingham, 1993b). These places can be very small areas or facilities that reflect and affect the activities of their users and may impact a specific criminal event (Loukaitou-Sideris, Liggett, Iseki, & Thurlow, 2001). The places may present two types of attractiveness to criminals: target attractiveness and spatial attractiveness(Rhodes & Conly, 2017). Target attractiveness refers to the presence of a certain level of victims at the places. Spatial attractiveness refers to the physical features of the places and their nearby environments that may facilitate crimes to occur unnoticed. Many empirical studies have proved that many business or public facilities people routinely visit can be crime attractors or generators, which provide more potential crime opportunities and increase crime in their vicinity, such as restaurants (Bernasco, Block, Rengert, Groff, & Eck, 2011), convenience stores (Askey, Taylor, Groff, & Fingerhut, 2018), department stores (Carroll & Weaver, 2017), neighborhood parks (Groff & McCord, 2012), stadiums (Kurland, Johnson, & Tilley, 2014), alcohol outlets (Day, Breetzke, Kingham, & Campbell, 2012) and so forth. For example, schools are generally found to have a significant positive correlation with crime incidents at block-level studies (Weisburd, Ready, & Lawton, 2012). In addition, scholars found an association between bus stops and street robberies (Liu, Lan, Eck, & Kang, 2020), and a correlation between road networks densities and property crime (Du & Law, 2016). In terms of theft crime, some scholars also found that motor vehicle theft was concentrated on facilities on commercial land (Kinney, Brantingham, Wuschke, Kirk, & Brantingham, 2008). Song et al. (2018) used official crime data from a large Chinese city to verify that theft rates were related to the presence of retail facilities that shape daily activities from the opportunity perspective (Song et al., 2018).

In the recent decades, researchers start to consider environmental factors when modeling crime, for example, risk terrain modeling (RTM) (Caplan & Kennedy, 2010), Bayesian model (Hu, Zhu, Duan, & Guo, 2018; Law, Quick, & Chan, 2014), and discrete choice model (Bernasco, 2010; Picasso & Cohen, 2019). Among them, RTM receives more attention. RTM tries to identify features in the environment which attract crime, and model how the existence of these features may create more crime opportunities. It starts by selecting and weighing environmental features that are spatially related to crime incidents in cross-section. Then a statistical model is used to calculate and plot places that have higher statistical possibilities of crime (so-called risk terrains) (Kennedy, Caplan, & Piza, 2011; Wheeler & Steenbeek, 2021). While RTM captures individual influence from each environmental feature, it fails to consider interactive effect among these features. In addition, the prediction accuracy of RTM may be low because it normally fails to consider temporal influences.

Unlike traditional crime prediction models and RTM, which need specific algorithms set by researchers, machine learning crime prediction models solely rely on computers' automotive analyses. Variables including historic crime events, various environmental features, and even time can all be incorporated in machine learning models. Users do not need to specify a particular algorithm, and the computer will decide the most suitable function to train itself. Thus, machine learning models tend to have better model fits than traditional regression models (Liu, Liu, Liao, et al., 2018; Yi, Yu, Zhuang, & Guo, 2019). With the rapid development of artificial intelligence (AI) technology in recent years, various machine learning crime prediction models emerge. The representative ones are: neural network model (Rummens et al., 2017), random forest model (Alves et al., 2018), and graph convolution model (Wang, Zheng, Yang, & Wang, 2020). However, even though they may improve prediction accuracy, none of them have systematically considered the spatiotemporal environment variables, nor did they explain the prediction process in a transparent manner.

In order to improve the transparency and interpretability in machine learning, two different methods are proposed: ante-hoc and post-hoc (Molnar, Casalicchio, & Bischl, 2020). The ante-hoc method uses a simple and interpretable structure to train the models (Alvarez-Melis & Jaakkola, 2018); while the post-hoc method, conversely, allows models to be trained as they normally would, and then consider models' interpretability after the training. Additionally, interpretability in the post-hoc method can be subdivided into global interpretability and local interpretability. Global interpretability aims to help the audience understand the overall logic and mechanism behind complex models (Guidotti et al., 2019); and local interpretability aims to explain the decision-making process and the basis of the machine learning model for each input sample (Baehrens et al., 2009).

The interpretability of the machine learning model is very important for crime prediction. it is important for researchers to know the machine learns in the right way. Model interpretation can help us understand why a machine learning model makes such a decision and what variable plays the most important role in the learning and decision process. At the same time, being able to interpret the machine learning model helps improve the credibility of the model and the transparency of the prediction results. It is certainly not practical for the police department to rely on the “black box” model to direct crime prevention and crime control strategies. Thus, interpreting the machine learning models can help researchers and practitioners check the unique contribution (positive or negative, significant or not) of each variable.

As stated earlier, an interpretable approach is needed to let the audience know what is going on in the “black box”, so that the state-of-the-art performance can earn trust from stakeholders and practitioners. However, understanding how the machine makes the decision is a challenge. Many complex models with high accuracy (e.g., deep learning model) do not make transparent decisions. While simple regression models usually cannot achieve a comparable prediction accuracy. This conflict forces a trade-off between accuracy and interpretability. Thus, to reduce the unexpected deviation and improve transparency, we need both a reliable crime prediction model and a transparent procedure that is understandable by scholars, practitioners, and stakeholders. Guided by environmental criminology theories, our study uses the XGBoost machine learning method with necessary crime and environment variables to predict crime, and then interpret predictions with the SHAP method. XGBoost predicts future crime with historical crime data and environmental factors. Then SHAP serves as a machine learning interpreter and interprets the prediction from both global and local perspectives to reveal the contribution of each variable. The combination of an accurate machine learning method and efficient interpreter is the unique contribution of this study to the literature.

2. Methods

2.1. XGBoost

To achieve high prediction accuracy and high interpretability, we select the XGBoost model to predict crime. XGBoost is a widely recognized tree machine learning model which balances accuracy, scalability, and efficiency well (Mousa, Bakhit, Osman, & Ishak, 2018). According to the decision rules in this tree model, the given samples are categorized, and the prediction is made by calculating the scores in the leaves after the cumulative classification (Chen & Guestrin, 2016). Supposing the model has k decision trees, the model's equation is:

$$\widehat{y}_i = \sum_{i=1}^{k} f_k(x_i), f_k \in \mathscr{F} \tag{1}$$

The objective function is as following:

$$obj(t) = \sum_{i=1}^{n} l(y_i, \widehat{y}_i) + \Omega(f(t)) \tag{2}$$

where $\Omega(f(t)) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2$

For Eq. (2), $l(y_i, \widehat{y}_i)$ is the loss function with the target $y_i$and prediction $\widehat{y}_i$. $\Omega(f(t))$ is the complexity of the entire tree and it is a regular term of the objective function. T is the total number of leaf nodes. γ is the penalty coefficient to control the number of leaf nodes to prevent overfitting. λ is the regularization coefficient. ω_j is the weight of leaf nodes. The number of leaf nodes (T) and the vector norm of weight (ω_j) jointly determine the size of the regularization term.

When a new decision tree is generated, the residual of the previous prediction needs to be fitted , f_t is added to minimize the loss function. $\widehat{y}_i^{\,t}$ is the prediction of the ith instance at the tth iteration.

$$\widehat{y}_i^{\,t} = \widehat{y}_i^{\,t-1} + f_t(x_i) \tag{3}$$

Then the objective function can be expressed as:

$$Obj(t) = \sum_{i=1}^{n} l\left(y_i, \widehat{y}_i^{\,t-1} + f_t(x_i)\right) + \Omega f(t) \tag{4}$$

After applying Taylor series expansion to the objective function, the final objective function can be obtained:

$$obj(t) \cong \sum_{j=1}^{t} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2 \right] + \gamma T + constant \tag{5}$$

Where $G_j = \sum_{i \in I_j} g_i, H_j = \sum_{i \in I_j} h_i$.

Here g_i is the first derivative of the objective function, h_i is the second derivative of the objective function, and I_j is defined as the set of samples on each leaf $I_j = \{i | q(x_i) = j_i\}$.

XGBoost is a second-order Taylor expansion of the loss function and adds a regular term to the loss function. It can calculate the optimal solution for the whole model, measure the decline of the loss function and the complexity of the model, avoid overfitting, and improve the solution efficiency of the model (Mousa et al., 2018). Additionally, XGBoost is not affected by multicollinearity, so we can keep all influential factors in the model, though some of them may correlate with each other (Parsa, Movahedi, Taghipour, Derrible, & Mohammadian, 2020).

2.2. Shapley additive explanation (SHAP)

The interpretability of the tree ensemble method is rather important but can be hard to achieve. In some machine learning methods, when the weight of one influential factor increases, the importance of this factor would decrease, which is confusing (Lundberg et al., 2018). Shapley additive explanation (SHAP), as a machine learning interpreter, can address such problems (Lundberg & Lee, 2017). SHAP was proposed by Shapley based on Game Theory in 1953 (Shapley, 1953). The goal of SHAP is to provide a measure of the importance of features in machine learning models. Its working principle is shown in the following equation. In a model , $\varnothing_i(v)$ represents the attribute value for each feature i:

$$\varnothing_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(s)\right) \tag{6}$$

Here, S is a subset of the features used in the model, N is the vector of feature values, and n is the number of features. $v(s)$ is the prediction for feature values in the set S. $\sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n-|S|-1)!}{n!}$ represents the weight, and ($v(S \cup \{i\}) - v(s)$) indicates the change value before and after adding the new feature i. By comparing the attribute value of each feature, the importance of the feature can be sorted. The contribution of each feature to the model output is assigned according to its marginal contribution in comparison to other features measured in SHAP values.

Lundberg and Lee have developed a practical Python package to calculate the SHAP of tree models such as GBoost, CatBoost, and XGBoost (Lundberg & Lee, 2017). SHAP has received recognition from many researchers in the field of transportation, traffic control, and energy demand management (Mihaita, Liu, Cai, & Rizoiu, 2019; Movahedi & Derrible, 2020; Molnar et al., 2020).

3. Study area and data

The study area is the XT Paichusuo of ZG City, a coastal city in Southeast China that has more than 10 million population and ranks at the top for economic development. Public theft is one of the most common crime types in this city and especially in the study area. The total area of XT Paichusuo is 7.42 km2, with a population of 400,000. Of the total population, 177,000 are registered residents with Hukou, while the rest are non-residents (domestic migrant workers). A Paichusuo is a basic unit for policing in China. Each can devise and carry out the optimal policing strategy for its jurisdiction. Therefore, Paichusuo is an ideal unit for crime research in China. The XT Paichusuo has one of the highest crime rates in the city, so the selection of XT Paichusuo as the study area has significant implications for policing and crime prevention. We retrieve public theft data with precise spatiotemporal information in our study area that happened during 2017–2020 from the Public Security Bureau of ZG City. In addition, we also collect the spatial locations of the surveillance cameras, police stations, and police substations in the study area from the police department. The unit of analysis is the 150 m × 150 m grid, and the entire study area is divided into 371 grids. The resolution of 150 m is based on the previous studies and the police's practical knowledge (Leigh, Jackson, & Dunnett, 2016; Block, 2000; Santitissadeekorn, Short and Lloyd, 2018). If the grid is too small, incidents will concentrate on only several grids, while the larger grid will reduce the spatial resolution (Rummens & Hardyns, 2021). Besides, 150x150 m^2 is typically regarded as the largest foot patrol area that a single police officer can cover in one petrol session, and targeted foot patrol can increase visibility and accessibility to audiences(Williams, 2016). Visible policing can both reduce fear of crime and increase public confidence (Ariel, Weinborn, & Sherman, 2016).

We followed routine activity theory and crime pattern theory to select variables, so all variables in our crime prediction model are theory-driven (Table 1). Ambient population can be used as an indicator of potential victims as one of the three elements of the routine activity theory (Malleson & Andresen, 2016). In addition to the residential population, the ambient population also contributes to crime within an area (Andresen, 2011). Different types of population data such as Landscan data, travel survey data, social media data, et al. have been used to measure the ambient population in previous studies of crime models (Felson & Boivin, 2015; He et al., 2020; Kurland et al., 2014; Malleson & Andresen, 2015). The ambient population data used in this

Table 1
Independent variable descriptions.

Variable category	Variable name	Source and meaning
1	Ambient population (16–24)	Population size at corresponding age groups in each grid
2	Ambient population (25–44)	
3	Ambient population (45–59)	
4	Ambient population (60–69)	
5	Camera	Number of Skynet cameras in each grid
6	Dist_Police	The distance from the grid's centroid to the nearest police station or police sub-station
7	Restaurant	The numbers of different Points of interest (POIs). The data is from the navigation data of Daodaotong Map company.
8	Bus station	
9	Department store	
10	Internet café	
11	Entertainment venue	
12	School	
13	Bank	
14	Hotel	
15	Convenience store	
16	Richness	Affluence index of each grid from Smart Steps data of China Unicom
17	Road_Lengh	Length of all road segments within the grid

paper come from China Unicom, which is one of the three telecom operators in China and has 300 million active users every day. As each user conducts user-BS (user and base station) communication every 7 min on average, the user's location data is recorded in the same time interval (Wu et al., 2020). According to the age classification of WHO, young is 25–44, middle age is 45–59, and old age is 60 and above (Dyussenbayev, 2017). Since the minimum age of theft offenders eligible for criminal punishment is 16 years in China, we also added a youth population variable of 16–24 years old. We extracted the four age groups from this dataset: 16–24, 25–44, 45–59, and 60–69 respectively. This dataset also provides the richness index, which shows the user's aggregated movement insights and can show society's overall wealth level by evaluating their property value, cell phone value, phone bill, and so forth. The company uses its patented algorithm to score each user on a scale of 1–8 based on their behavioral profile attributes, and then calculates the average of this score for users within a grid to create an index of affluence of that grid. According to Unicom, the behavioral characteristics attributes include the price of housing in the subscriber's neighborhood, the price of cell phone terminal, the number of places to stay for entertainment, the amount of travel in foreign cities, the mode of travel, phone bills and other multi-source data indicators. This richness index ranges from 1 to 8, and a higher value indicates more assets. The accessibility variable is measured as the total length of the road segments in each grid. The numbers of restaurants, bus stations, department stores, Internet cafés, entertainment venues, schools, banks, hotels, and convenience stores are also included as they may act as crime attractors/generators and influence crime opportunities (Brantingham & Brantingham, 1993a; Lan et al., 2021).

The crime type studied in this paper is public theft, which refers to theft that happened in public places. Pickpocketing, thefts from malls and convenience stores, theft of electric vehicles, bicycles, and motorcycles are all considered as public theft (Liu et al., 2017). From 2017 to 2020, in our study area, public theft accounted for more than 40% of all crime incidents. Our dependent variable data is binary, showing the presence or not of public theft in the grid. This is because more than 90% of grids have no cases in the two-week statistical cycle, and no grid has more than 9 thefts. The time scale used in this paper is two weeks, which is recommended by local police officers. This is a practical time slot that police officers would use to adjust their strategies.

4. Results

4.1. *The result of the* crime prediction model

In total, 371 grids (150 m × 150 m) are used for crime prediction models. We divide historical public thefts into biweekly cycles and receive 78 cycles during 2017–2019, so the sample size of the data set is 28,938 (371 × 78 = 28,938). In our study area, public thefts are concentrated, more than 93% (27,005/28,938) grids have no theft problem. We use Python to conduct crime prediction with the XGBoost machine learning algorithm. First, we divide all samples into the training group (75% of all samples) and the verification group (25% of all samples). Then, the XGBoost model is fit, and the grid search method is used to adjust the parameters of the model as an optimization (Putatunda & Rama, 2018). Finally, the best model is automatically decided based on the evaluation index by computer. Cross-validation is used to evaluate the performance of various combinations of parameters, and the best combination is selected as the model parameters.

The accuracy of the training model is 0.91, suggesting that 91% of the grids are correctly predicted. The accuracy of the verification set is 0.89, and the ROC (Receiver Operating Characteristic)score is 0.586. In addition, we also compare the XGBoost model with other popular machine learning models such as logistic regression, decision tree, and random forest; and the XGBoost model clearly shows the best model fit (Fig. 1).

4.2. *Global interpretability*

SHAP tree explainer is used to explain the model both globally and locally. The SHAP value of a feature (variable) shows its contribution to the output, and such a value is weighted and summed over all possible feature value combinations. Global interpretability refers to the ability to interpret model decisions based on conditional interactions between dependent and independent features in the entire dataset. It shows the overall influence of model features and how each algorithm component (e.g., weight, structure, and other parameters) help machine make decisions. The SHAP value, $\varnothing_i(f,x)$, is an allocation of credit among the various features in the feature set to explain a prediction $f(x)$. In our crime prediction model, f is the XGBoost model results and x is the variable. The mean value of all absolute SHAP values ($\widehat{\varnothing\_i(f,x)}$) shows the global interpretability of the crime prediction model.

Fig. 2 ranks the mean absolute SHAP value of each variable from high to low. Variables are sorted according to their impact, and the

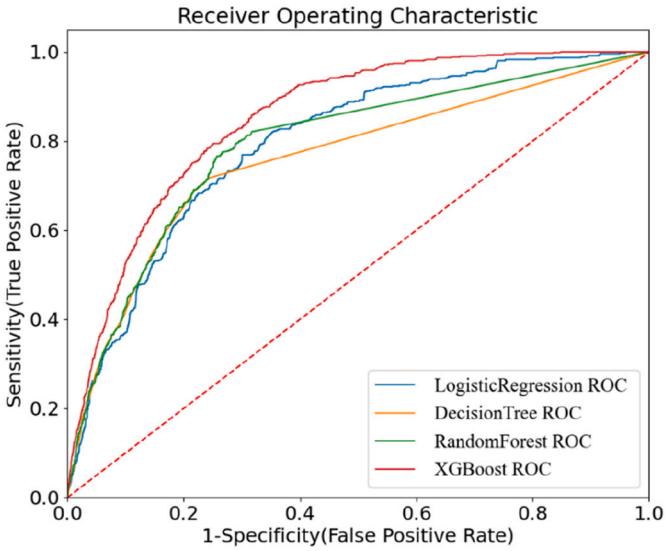


Fig. 1. Comparison of ROC curves of different models.

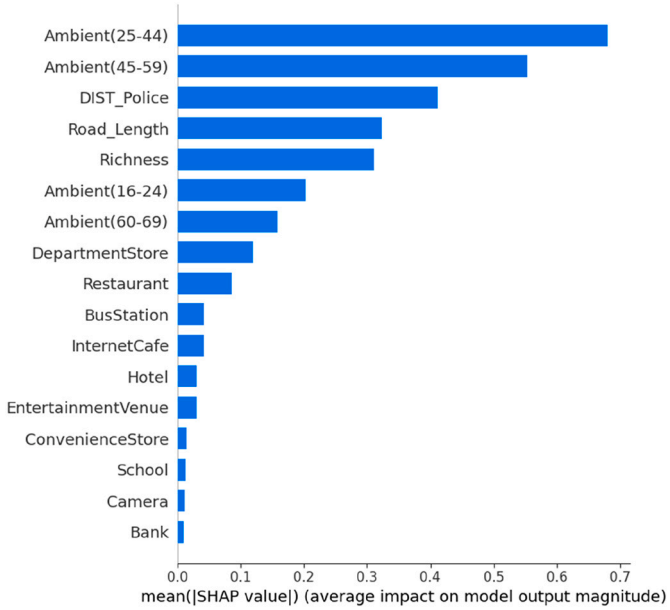


Fig. 2. Ranking of the absolute value of SHAP value of all features.

variables at the top have the greatest impact. The top two variables are *ambient (25–44)* and *ambient (45–59)*. This means the ambient population at ages of 25–59 and 45–59 have the best global interpretability to the crime prediction in the study area.

To better understand the impact of each variable on the model, we plot the SHAP value of each variable at each grid (Fig. 3). The x-axis shows the impact of variables on the outcome. Positive SHAP values indicate a positive relationship between the independent variable and dependent variable, while negative values indicate negative relationships.The attribute values of the samples are indicated by the colors of the dots (red color indicates high attribute value and blue color indicates low attribute value) Fig. 3 shows that the ambient population at ages of 25–44 (*ambient(25–44)*) has the greatest impact on crime. This means grids with a larger population of 25–44 are more prone to have public

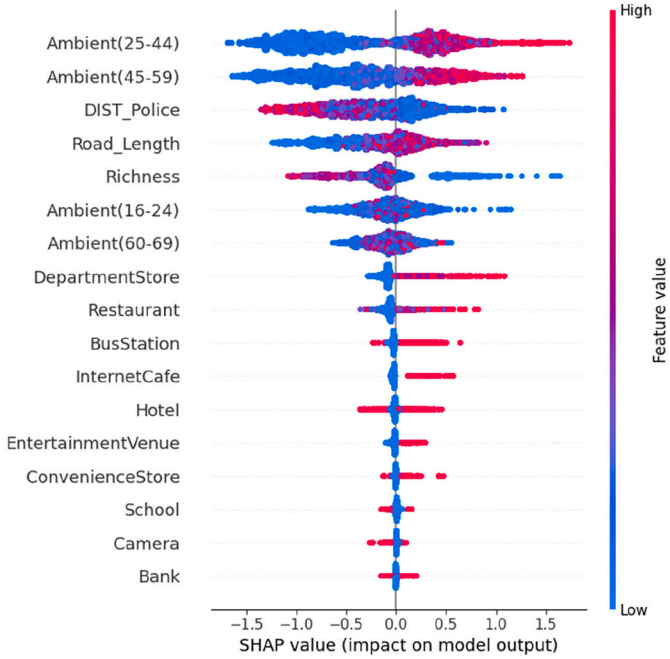


Fig. 3. Distribution of SHAP values of all samples.

theft problems than other grids. The variable of the ambient population at ages of 45–59 (*ambient(45–59)*) has the second-highest impact on the model. For the variable *DIST_ Police*, the red dots are mostly on the left side of the y-axis, which indicates that when the feature value of the variable *DIST_ Police* is large, its SHAP value is negative. This means the risk of public theft is high near the location of police stations/sub-stations. This may seem controversial, but it is actually reasonable, we will explain this in detail in the later paragraph. Other variables, such as the number of department stores, restaurants, bus stations, and internet cafes, have less impact on the model. However, when their values are large, their SHAP values are positive.

To further study the interrelationships between crime and each variable, we plot variables with variable values on the x-axis and its SHAP value on the y-axis (Fig. 4). Fig. 4(a) and Fig. 4(b) illustrate the impact of the ambient population (25–44) and ambient population (45–59) on model output. Larger variable values generally relate to higher SHAP values, which indicates that they are positively correlated with public theft. Thus, those grids with a large proportion of the population of 25–59 are more likely to be the target area of public theft. These ambient populations characterize the potential victims. The increase of potential victims increases the risk of crime to a great extent. On the contrary, the value of *DIST_Police* and SHAP value are negatively related (Fig. 4(c)). This means that grids closer to the police station/sub-station tend to have a higher theft risk. This is because, in our study area, police stations and police sub-stations are generally located in an area where there are a more dynamic population and a higher crime risk. Fig. 4(d) shows that the high value of the variable of *ambient(25–44)* is concentrated in the area with a small variable value of *DIST_Police* (the red points are on the left part of the figure), which means that grids with more 25–44 aged population and closer to police stations/sub-stations tend to experience more public theft. Plots such as Fig. 4(d) helps reveal interactive effects between the different variables.

While the machine learning model does not explicitly specify the interaction between the explanatory variables, it can assess the individual contribution of a variable in comparison with the others, in a way similar to the controls used in a regular regression.

To test the consistency between the machine learning model and traditional regression model, we performed logistic regression modeling and obtained the results in Table 2. Among the statistically significant variables, *Ambient (25–44)*, *Ambient (45–59)*, *Ambient (60–69)*, *Restaurant*, *DepartmentStore*, *InternetCafe* and *ConvenienceStore*have an odds ratio greater than 1 and a z-value greater than 0 which means that these variables can increase the risk of crime. Observing Fig. 3, the SHAP model also has the same results. While *Ambient (16–24)*, *DIST_Police*, *School* and *Richness* have an odds radio less than 1 and a z-value less than 0, implying that the relationship between crime occurrence of these variables is negative. These odds ratios are consistent with those presented by the SHAP model.

4.3. Local interpretability

The local interpretability of the model suggests the prediction contribution of every single grid. That can be achieved by analyzing the contribution of each dimensional feature of each individual sample to the model outputs (Lundberg & Lee, 2017). Through the local variation of different feature' SHAP values, we can get a sense of the local microenvironment in each grid. The goal of SHAP value calculation is to explain the result of the machine judgment by calculating the contribution of each feature when predicting. The sum of the SHAP values of each feature in a sample plus the baseline value equals the predicted value of the sample. The local accuracy, also known as additivity, can be calculated as follows:

$$f(x) = \varnothing_0(f,x) + \sum_{i=1}^{V} \varnothing_i(f,x) \tag{7}$$

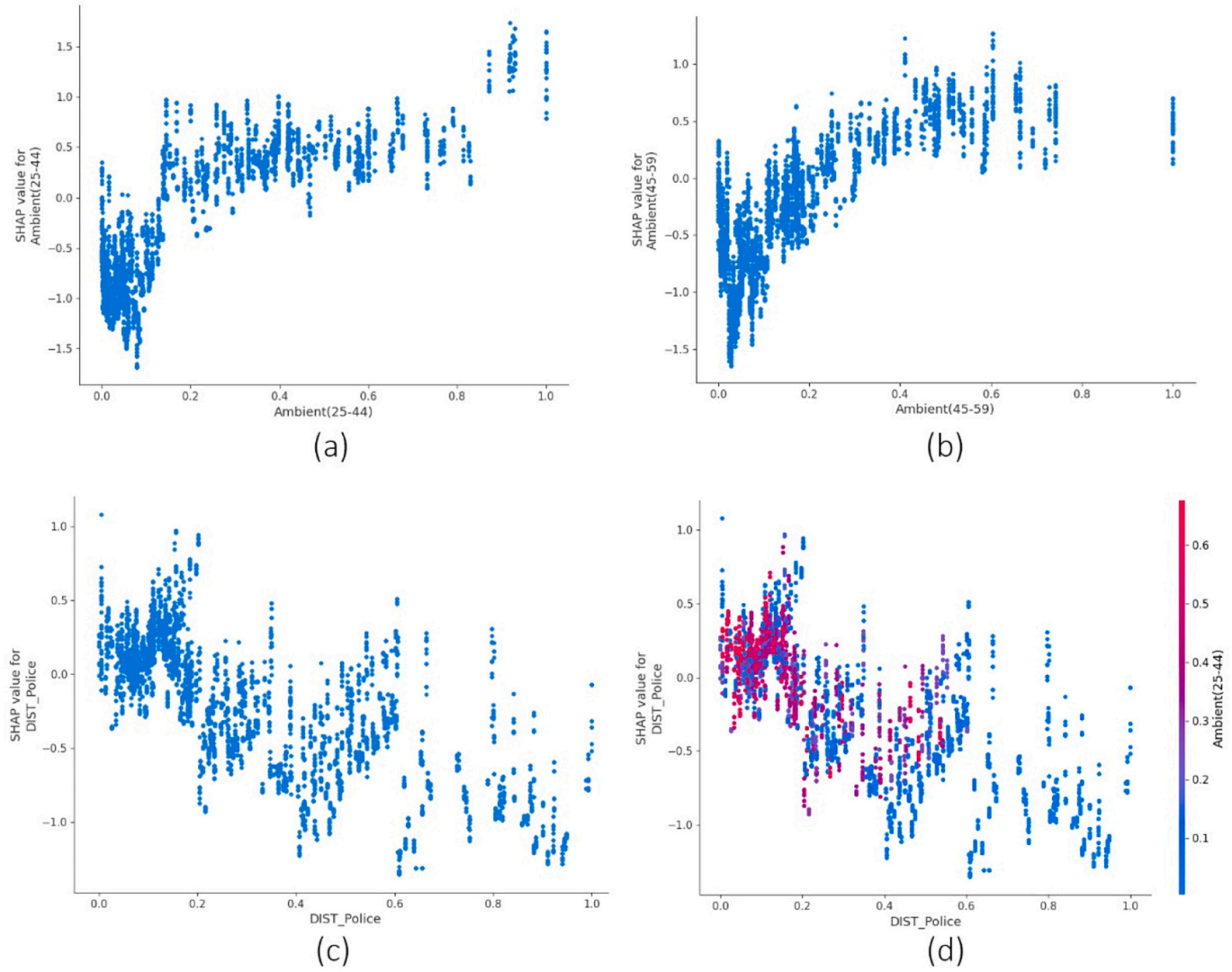


Fig. 4. Dependency plot of variables and SHAP value.

where $\varnothing_0(f,x) = E\ [f\ (\ x)]$ Represents the expected crime risk of the model over the training dataset and V is the number of inputs.

Additionally, the SHAP value has monotonicity, which is also known as consistency. Specifically, if a feature is more important in one model than other models, no matter what other features are also present in other models, the importance attributed to this very feature should also be higher (Lundberg et al., 2018). For the crime prediction model, the feature importance of one grid does not necessarily mean causality. However, such feature importance can help direct police officers when making crime prevention strategies. We randomly select two grids and show the SHAP value of each feature in Fig. 5. Fig 5(a) and 5(b) show two different grids, and they share the same base value ($\varnothing_0 = -2.41$), indicating mean value of target variable of all samples. Blue color indicates the negative influence of that feature, and red color indicates the positive influence. As shown in Fig. 5(a), features with negative SHAP values are ambient *population(24–44)*, *ambient population(45–59)*, and *ambient population(16–24)*; while variables with positive SHAP values are *Bus Station* , *DIST_ Police* and *Road_ Length*. The total SHAP value of the sample is −3.38, which is smaller than the base value. In Fig. 5(b), the variable with negative SHAP value is *Richness*, and the variables with positive SHAP values are *ambient population(45–59), ambient population (60–69), ambient population(24–44), road_ Length, DIST_ Police*, and *Entertainment Venue*, etc. These two examples show that the direction of a feature's local influence on crime can be different from that of the global model, and the local influences vary across the grids.

Fig. 6 shows the local SHAP values of four different variables in the training model by the grid. Fig. 6(a) and 6(b) show variables of the *ambient population(25–44)* and *ambient population(45–59)*, while Fig. 6 (c) and Fig. 6(d) show *department stores* and *restaurants*, which are crime attractors and generators. These graduated symbol maps can suggest crime hot spots that are dominated by that particular variable. The distributions of local SHAP values of different variables are also different, which means that different variables have different impacts on the model at different grids. These maps, which show unique contributions from every single variable, can provide meaningful information to the police to create tailored strategies for crime prevention. This is the unique contribution of this interpretable machine learning crime prediction method.

5. Discussion and conclusions

Existing crime prediction models using the machine learning method tend to act as the “black box”. It is almost impossible to understand what happens in the “black box”. The lack of interpretability may undermine people's confidence in these crime prediction models. This study has overcome this limitation, and the resulting models not only increased predictability but also brought interpretability, similar to those of

Table 2
The result of logistic regression.

Y	Odds ratio	St.Err.	z	P > z	[95% Conf	Interval]	Sig
Ambient (16–24)	0.999	0.000	−9.53	0.000	0.998	0.999	***
Ambient (25–44)	1.002	0.000	17.13	0.000	1.002	1.003	***
Ambient (45–59)	1.009	0.001	12.16	0.000	1.008	1.011	***
Ambient (60–69)	1.015	0.004	3.43	0.001	1.006	1.024	***
Camera	1.01	0.026	0.38	0.701	0.961	1.061	
DIST_Police	0.999	0.000	−11.51	0.000	0.999	0.999	***
Restaurant	1.041	0.018	2.36	0.018	1.007	1.076	**
BusStation	1.033	0.045	0.74	0.462	0.948	1.125	
DepartmentStore	1.059	0.021	2.86	0.004	1.018	1.102	***
InternetCafe	1.748	0.152	6.42	0.000	1.474	2.074	***
EntertainmentVenue	1	0.061	−0.01	0.996	0.887	1.126	
School	0.843	0.04	−3.57	0.000	0.768	0.926	***
Bank	1.063	0.07	0.93	0.352	0.935	1.209	
Hotel	0.989	0.036	−0.31	0.756	0.92	1.063	
ConvenienceStore	1.413	0.08	6.10	0.000	1.265	1.579	***
Richness	0.866	0.013	−9.24	0.000	0.84	0.893	***
Road_Length	0.999	0.000	−1.67	0.095	0.999	1	*
Constant	0.087	0.007	−29.62	0.000	0.074	0.102	***
Mean dependent var	0.099	SD dependent var.	0.299				
Pseudo r-squared	0.162	Number of obs	28,938				
Chi-square	2798.891	Prob > chi2	0.000				
Akaike crit. (AIC)	14,510.992	Bayesian crit. (BIC)	14,658.463				

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

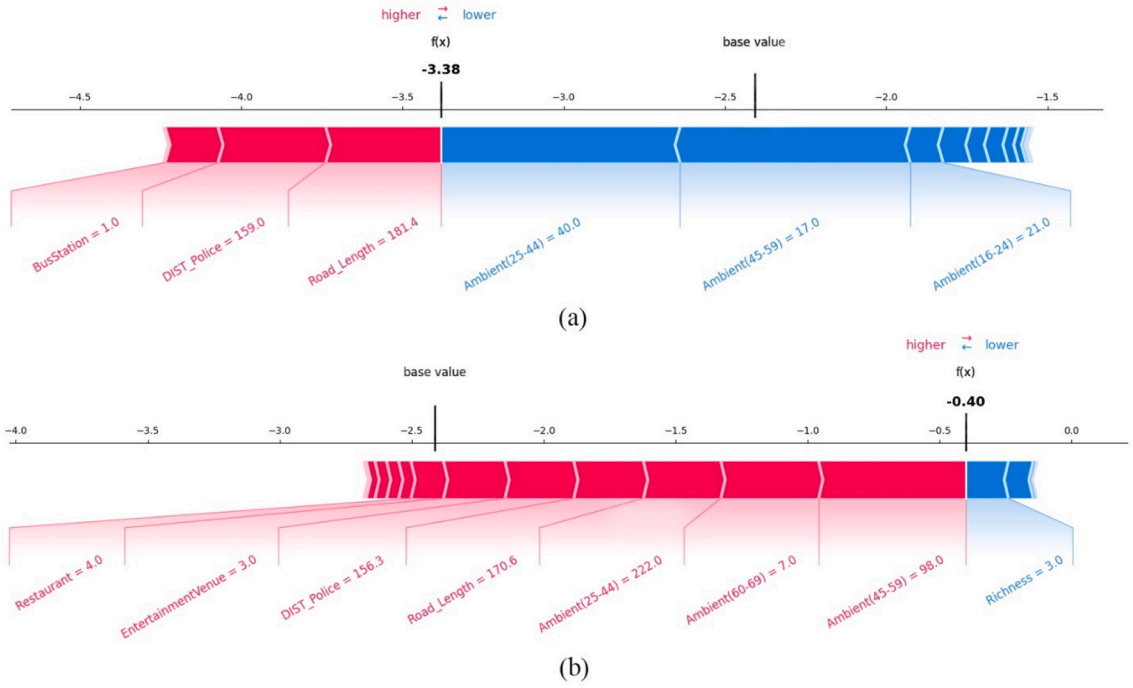


Fig. 5. Local SHAP values of the features at two randomly selected grid locations.

traditional regression models.

Our study proposes an interpretable machine learning crime prediction model by using XGBoost and SHAP. By increasing the transparency and interpretability of the machine learning crime prediction model, our approach can provide practical insights to practitioners. In terms of practical application, the knowledge obtained from the interpretability analysis of the model can provide insights to help police officers formulate data-driven policies and tailored crime prevention strategies. However, we do acknowledge this model is only tested in one city. Though it shows promising results, more tests are needed to further test its general applicability for other crime types in other study areas. The interpretability analysis of other machine learning and deep learning crime prediction models may also be noteworthy to conduct. Ideally, if we had Unicom Smart Footprint data for the entire study period, we could match ambient population perfectly to the crime during the same period. However, we only have the data for seven consecutive days. The seven-day average is used to represent daily ambient population. While this is not ideal, it is still reasonably representative for our models, which are based a two-week interval, since ambient population is not expected to change significantly from one week to another. Further, the granularity of the model can be refined. The time interval of crime statistics for the grid is two weeks in this study. The effect of each variable on crime in smaller time intervals can be explored in the future. Each day can also be divided into morning,

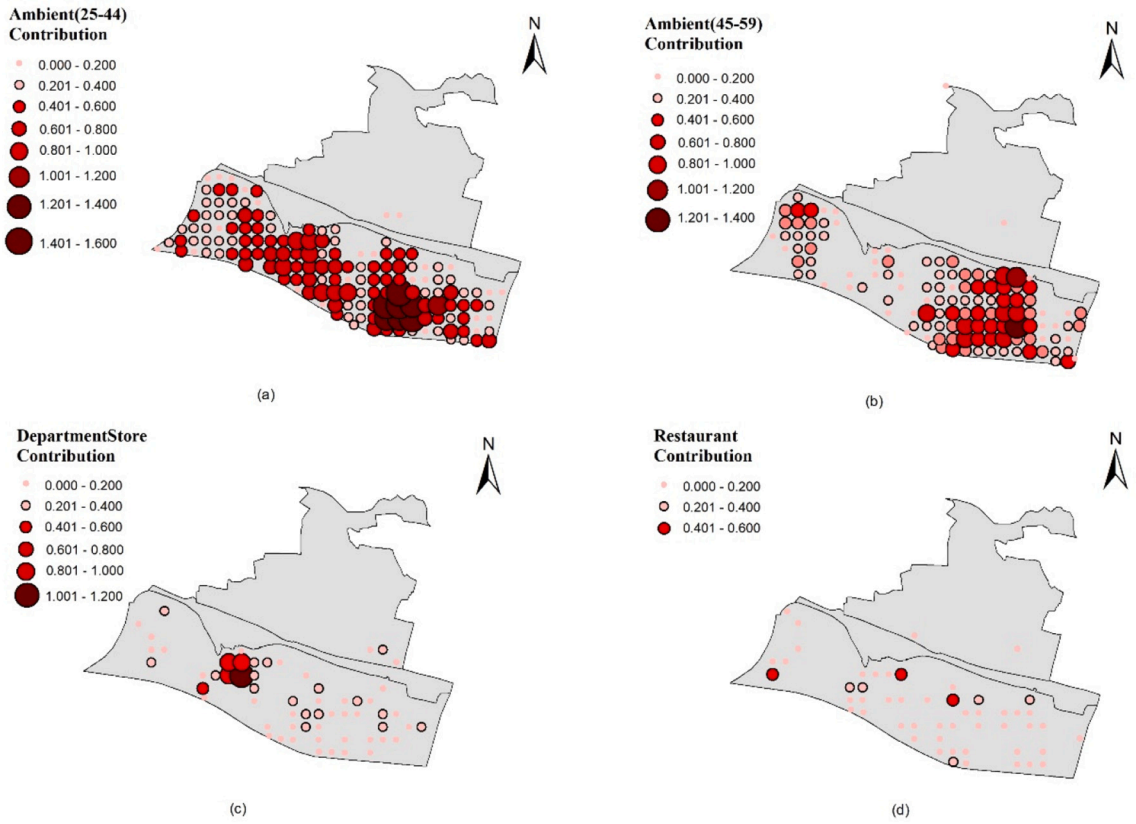


Fig. 6. Contribution of different variables in the model over space (Dec. 18 to Dec. 31, 2019).

midday and evening to explore the effect of various influencing factors on crime at different times of a day. In addition, the ambient population can be divided to more age groups, and the gender ratio and migrant population ratio of the ambient population could also be inferred.

In conclusion, our study specifically tackles machine learning models' problem of interpretability and creates a transparent and interpretable crime prediction model with machine learning methods. First, we identified the XGBoost out of several machine learning models as the best model for crime prediction. Based on the SHAP value, we ranked the contributions of all variables and found that the young population aged 25–44 contributed most to public theft in the study area. Further, we calibrated local XGBoost models and revealed the precise spatial variations of each variable in the study area. These findings help the local police force create more targeted policing strategies for crime reduction. For example, the police department can be advised to increase patrols in areas with high SHAP value of ambient population variables for the corresponding age group to enhance supervision.

Author Statement

The authors declare no conflict of interest. We sincerely appreciate anonymous reviewers for their advice to improve the quality of this study.

Data availability

The authors do not have permission to share data.

Acknowledgements

This work was supported by the Natural Science Foundation of Guangdong Province (grant number 2019A1515011065), National Natural Science Foundation of China (grant numbers 42001171, 41901177, 42171218, 42071184) and the Innovation Research Grant for the Postgraduates of Guangzhou University [grant number 2020GDJC-D02]

References

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 7786–7795).

Alves, L. G. A., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications, 505*, 435–443.

Andresen, M. A. (2011). The ambient population and crime analysis. *The Professional Geographer, 63*(2), 193–212.

Apicella, A., Isgrò, F., Prevete, R., & Tamburrini, G. (2020). Middle-level features for the explanation of classification systems by sparse dictionary methods. *International Journal of Neural Systems, 30*(08), 2050040.

Ariel, B., Weinborn, C., & Sherman, L. W. (2016). "Soft" policing at hot spots—Do police community support officers work? A randomized controlled trial. *Journal of Experimental Criminology, 12*(3), 277–317.

Askey, A. P., Taylor, R., Groff, E., & Fingerhut, A. (2018). Fast food restaurants and convenience stores: Using sales volume to explain crime patterns in Seattle. *Crime & Delinquency, 64*(14), 1836–1857.

Baehrens, D., et al. (2009). How to explain individual classification decisions. *Journal of Machine Learning Research, 11*, 1803–1831.

Bernasco, W. (2010). Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places. *Journal of Quantitative Criminology, 26*(1), 113–138.

Bernasco, W., Block, R., Rengert, G., Groff, E., & Eck, J. (2011). Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. *The Journal of Research in Crime and Delinquency, 48*(1), 33–57.

Block, R. (2000). Gang Activity and Overall Levels of Crime: A New Mapping Tool for Defining Areas of Gang Activity Using Police Records. *Journal of quantitative criminology, 16*(3), 369–383.

Brantingham, P., & Brantingham, P. (1993a). *Environment, routine, and situation: Toward a pattern theory of crime: Advances in criminological theory* (pp. 259–294). New Brunswick: Clarke and Marcus Felson.

Brantingham, P. L., & Brantingham, P. J. (1993b). Nodes, paths and edges: Considerations on the complexity of crime and the physical environment. *Journal of Environmental Psychology, 13*(1), 3–28.

Caplan, J. M., & Kennedy, L. W. (2010). *Risk terrain modeling manual: Theoretical framework and technical steps of spatial risk assessment for crime analysis*. Rutgers Center on Public Security.

Carroll, J., & Weaver, F. (2017). Shoplifters' perceptions of crime opportunities: A process-tracing study. In *The Reasoning Criminal* (pp. 19–38). Routledge.

Chainey, S., & Ratcliffe, J. (2013). *Identifying crime hotspots* (pp. 145–182). Chichester, West Sussex, England: John Wiley & Sons, Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *ACM*, 785–794.

Choy, G., et al. (2018). Current applications and future impact of machine learning in radiology. *Radiology, 288*(2), 318–328.

Cohen, L. E. (1981). Modeling crime trends: A criminal opportunity perspective. *Journal of Research in Crime and Delinquency, 18*(1), 138–164.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608.

Cornish, D., & Clarke, R. (1987). Understanding crime displacement: An application of rational choice theory: Routine activity. *Rational Choice and their Variants. Criminology, 25*(4), 933–947.

Day, P., Breetzke, G., Kingham, S., & Campbell, M. (2012). Close proximity to alcohol outlets is associated with increased serious violent crime in New Zealand. *Australian and New Zealand Journal of Public Health, 36*(1), 48–54.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. *IEEE*, 8599–8603.

Du, Y., & Law, J. (2016). How do vegetation density and transportation network density affect crime across an urban central-peripheral gradient? A case study in Kitchener—Waterloo, Ontario. *ISPRS International Journal of Geo-Information, 5*(7), 118.

Dyussenbayev, A. (2017). View of age periods of human life. *Advances in Social Sciences Research Journal, 4*(6).

Farrell, G., & Pease, K. (1993). *Once Bitten*. Twice Bitten: Repeat Victimisation and Its Implications for Crime Prevention.

Felson, M., & Boivin, R. (2015). Daily crime flows within a city. *Crime Science, 4*(1), 31.

Groff, E., & McCord, E. S. (2012). The role of neighborhood parks as crime generators. *Security Journal, 25*(1), 1–24.

Guidotti, R., et al. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

He, L., et al. (2020). Ambient population and larceny-theft: A spatial analysis using mobile phone data. *ISPRS International Journal of Geo-Information, 9*(6), 342.

Hoermann, S., Bach, M., & Dietmayer, K. (2018). Dynamic occupancy grid prediction for urban autonomous driving: A deep learning approach with fully automatic labeling. In *2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia*.

Hu, T., Zhu, X., Duan, L., & Guo, W. (2018). Urban crime prediction based on spatio-temporal Bayesian model. *PLoS One, 13*(10), Article e0206215.

Kalinic, M., & Krisp, J. (2018). Kernel density estimation (KDE) vs. hot-spot analysis - detecting criminal hot spots in the city of San Francisco, Agile 2018. In *21st Conference on Geo-information Science, Lund, Sweden*.

Kennedy, L. W., Caplan, J. M., & Piza, E. (2011). Risk clusters, hotspots, and spatial intelligence: Risk terrain modeling as an algorithm for police resource allocation strategies. *Journal of Quantitative Criminology, 27*(3), 339–362.

Kinney, J. B., Brantingham, P. L., Wuschke, K., Kirk, M. G., & Brantingham, P. J. (2008). Crime attractors, generators and detractors: Land use and urban crime opportunities. *Built Environment (London. 1978), 34*(1), 62–74.

Kurland, J., Johnson, S. D., & Tilley, N. (2014). Offenses around stadiums: A natural experiment on crime attraction and generation. *The Journal of Research in Crime and Delinquency, 51*(1), 5–28.

Lan, M., Liu, L., & Eck, J. E. (2021). A spatial analytical approach to assess the impact of a casino on crime: An example of JACK casino in downtown Cincinnati. *Cities, 111*, Article 103003.

Law, J., Quick, M., & Chan, P. (2014). Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level. *Journal of Quantitative Criminology, 30*(1), 57–78.

Leigh, J., Jackson, L., & Dunnett, S. (2016). Police officer dynamic positioning for incident response and community presence. In *Proceedings of the 5th International Conference on Operations Research and Enterprise Systems (ICORES 2016)* (pp. 261–270).

Liu, L., Du, F., Xiao, L., Song, G., Liu, K., & Jiang, C. (2017). The Density of Various Road Types and Larceny Pate: LARCAn Empirical Analysis of ZG City. *Human Geography, 32*(06), 32–46.

Liu, L., Lan, M., Eck, J. E., & Kang, E. L. (2020). Assessing the effects of bus stop relocation on street robbery. *Computers, Environment and Urban Systems, 80*, Article 101455.

Liu, L., Liu, W. J., Liao, W., et al. (2018). Comparison of random forest algorithm and space-time kernel density mapping for crime hotspot prediction. *Progress in Geography, 37*(6), 761–771.

Loukaitou-Sideris, A., Liggett, R., Iseki, H., & Thurlow, W. (2001). Measuring the effects of built environment on bus stop crime. *Environment and Planning B: Planning and Design, 28*(2), 255–280.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. In *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*.

Lundberg, S. M., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749–760.

Malleson, N., & Andresen, M. A. (2015). The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science, 42*(2), 112–121.

Malleson, N., & Andresen, M. A. (2016). Exploring the impact of ambient population measures on London crime hotspots. *Journal of Criminal Justice, 46*, 52–63.

Mihaita, A., Liu, Z., Cai, C., & Rizoiu, M. (2019). *Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning – a brief history, state-of-the-art and challenges. *arXiv e-prints*, 417–431. arXiv:2010.09337.

Mousa, S. R., Bakhit, P. R., Osman, O. A., & Ishak, S. (2018). A comparative analysis of tree-based ensemble methods for detecting imminent lane change maneuvers in connected vehicle environments. *Transportation Research Record: Journal of the Transportation Research Board, 2672*(42), 268–279.

Movahedi, A., & Derrible, S. (2020). *Interrelated patterns of electricity, gas, and water consumption in large-scale buildings*. https://engrxiv.org/preprint/view/792.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, Article 105405.

Picasso, E. A. E. U., & Cohen, M. A. A. (2019). Valuing the public's demand for crime prevention programs: A discrete choice experiment. *Journal of Experimental Criminology, 4*, 529–550.

Putatunda, S., & Rama, K. (2018). A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. *ACM*, 6–10.

Rhodes, W. M., & Conly, C. (2017). Crime and mobility: An empirical study principles of geographical offender profiling. *Routledge*, 143–164.

Rummens, A., & Hardyns, W. (2021). The effect of spatiotemporal resolution on predictive policing model performance. *International Journal of Forecasting, 37*(1), 125–133.

Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context. *Applied Geography, 86*, 255–261.

Santitissadeekorn, N., Short, M. B., & Lloyd, D. J. B. (2018). Sequential data assimilation for 1D self-exciting processes with application to urban crime data. *Computational Statistics & Data Analysis, 128*, 163–183.

Shapley, L. S. (1953). *Contributions to the Theory of Games II*. Princeton University Press.

Sherman, L., Gartin, P., & Buerger, M. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology, 27*, 27–56.

Song, G., et al. (2018). Theft from the person in urban China: Assessing the diurnal effects of opportunity and social ecology. *Habitat International, 78*, 13–20.

Stummvoll, G. (2009). Environmental criminology and crime analysis. *Crime Prevention and Community Safety, 11*(2), 144–146.

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708).

Townsley, M. (2003). Infectious burglaries. A test of the near repeat hypothesis. *British Journal of Criminology, 43*(3), 615–633.

Wang, M., Zheng, K., Yang, Y., & Wang, X. (2020). An explainable machine learning framework for intrusion detection systems. *IEEE Access, 8*, 73127–73141.

Wang, Y., Ge, L., Li, S., & Chang, F. (2020). In G. Dobbie, U. Frank, G. Kappel, S. W. Liddle, & H. C. Mayr (Eds.), *Deep temporal multi-graph convolutional network for crime prediction* (pp. 525–538). Cham: Springer International Publishing.

Weisburd, D., Lawton, B., & Ready, J. (2012). *Staking Out the Next Generation of Studies of the Criminology of Place* (pp. 236–243).

Weisburd, D., Ready, J., & Lawton, B. (2012). *Staking Out the Next Generation of Studies of the Criminology of Place* (pp. 236–243). New York: Oxford University Press.

Wheeler, A. P., & Steenbeek, W. (2021). Mapping the risk terrain for crime using machine learning. *Journal of Quantitative Criminology., 37*(2), 445–480.

Wilcox, P., & Eck, J. E. (2011). Criminology of the unpopular: Implications for policy aimed at payday lending facilities. *Criminology & Public Policy, 10*, 473.

Williams, S. (2016). *Do visits or time spent in hot spots patrol matter most? A randomised control trial in the west midlands police*. M. St Thesis in Applied Criminology and Police Management. Cambridge: University of Cambridge.

Wu, Y., et al. (2020). Comparison of the spatiotemporal mobility patterns among typical subgroups of the actual population with mobile phone data: A case study of Beijing. *Cities, 100*, Article 102670.

Yi, F., Yu, Z., Zhuang, F., & Guo, B. (2019). Neural network based continuous conditional random field for fine-grained crime prediction. In *Twenty-Eighth International Joint Conference on Artificial Intelligence {IJCAI-19}* (pp. 4157–4163).

Zhang, X., Liu, L., Xiao, L., & Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access, 8*, 181302–181310.