ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

Research on Crime Occurrence Prediction Using Machine Learning Methods—Considering Four Types of Crime in Chicago

Yining Xiong

School of International Police Studies, People's Public Security University of China, Beijing, 100038, China

Abstract: *The establishment of crime prediction models using big data has become a key part of current police work. This paper combines binary classification with SVM, Random Forest, XGBoost, and GBDT methods based on the data of burglary, battery, assault, and criminal damage in Chicago from 2015 to 2020, and compares the prediction results. In this experiment, GBDT and XGBoost presented relatively stable and excellent data, and the scores of F1-score were 0.69 and 0.71 respectively, with high scores. It means that the model has strong generalization ability, can better adapt to different data distributions, and can be used more in the future to provide decision support for local police work.*

Keywords: *Machine learning, Crime occurrence prediction, XGBoost, and GBDT*

1. Introduction

At present, in the context of big data, crime prediction technology builds models by analyzing historical crime data, socio-economic indicators, geographic information and other multidimensional data, and uses algorithms such as machine learning to predict the time, spatio-temporal and type of crime. The application of this technology has transformed police work from a passive response to active prevention. Through prediction results, the police can optimize the deployment of police forces, patrol and monitor high-risk areas in advance, effectively reduce crime rates, and improve the efficiency and effectiveness of police work.

Chicago, as a metropolis, has distinct characteristics and trends in crime. The crime rate in Chicago is relatively high, especially for violent and property crimes. Specifically, the violent crime rate in Chicago is one of the highest in all communities in the United States, including rape, murder and non-negligent manslaughter, armed robbery, burglary, and serious assault.

In accordance with the needs of this article, we introduce four types of crimes. Burglary refers to the illegal entry into any building or structure with the intent to commit theft or other criminal activities. In Chicago, burglary is a common type of property crime, and according to FBI crime data, the burglary rate in Chicago is 28 per thousand residents, meaning that one out of every 36 people may become a victim. (data source: https://www.neighborhoodscout.com/il/chicago/crime). Battery refers to the act of illegally using violence or threatening to use violence to cause harm to others. Battery in Chicago crime statistics is classified as an assault crime, including various forms of violent behavior, such as threats and beatings. Assault refers to the act of threatening to illegally cause physical harm to others, even without actual physical contact. Assault crimes hold an important position in Chicago crime statistics, including various forms of threats and assault behaviors. Criminal Damage refers to the intentional destruction or damage of another person’s property, usually involving malicious or reckless behavior. Criminal damage in Chicago crime statistics is classified as property crime, involving acts of intentionally destroying or damaging another person’s property.

Therefore, this study will use the data of burglary, battery, assault, and criminal damage from 2015 to 2020 to analyze the crime situation in Chicago, build a basic model for crime prediction, provide help for the development of police work, and lay the foundation for the construction of more in-depth crime prediction models in the future.

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

2. Research Review

Machine learning is one of the main methods for crime prediction at present. In current research, high-frequency models include Random Forest models, spatiotemporal neural network models, Hawkes process models, etc. For example, in predicting theft cases, the Long Short-Term Memory model (LSTM) appears frequently. Using time series models can analyze the statistical patterns in historical time series data and further predict crime risks. Timely and effective predictions can provide decision support for the adjustment of community safety prevention measures, patrol police calculation and deployment, and other specific business work [1-2].

The prediction of crime in China started relatively late. In the early days, domestic crime prediction research was mostly empirical prediction. As time progressed, it was not until the 21st century that mathematical analysis methods began to be used [3]. Han Yishi, Fan Yingsheng, Li Guojun, etc. [4] used the ARIMA model to predict the number of communication network fraud crimes in Zhejiang Province from January to June 2015, and the average error after taking the absolute value was about 9.2%, with good prediction results. Hou Miaomiao [5] and other scholars used the SARIMA time series prediction model to detect cycles and predict trends for robbery, snatching, and general assault crimes in a large city in northern China. The prediction results are good, with an average prediction relative mean squared error (PRMSE) of 11.95% and an average absolute percentage error (MAPE) of 10.92%. Liu Xiaojuan et al. [6] used the grey system theory analysis method to dynamically analyze criminal cases from 1990 to 1995, and the model accuracy P0 can reach 93.5%, with relatively ideal test results. Although these models have achieved good results, with the advent of the big data era, the amount and types of data have increased greatly. To meet the current needs for crime prediction and the need for accuracy, professional researchers in crime prediction have included more machine learning algorithms. For example, Alves et al [7]. trained a random forest regression model using 13 urban indicators to predict the monthly homicide rates across Brazil, achieving an accuracy rate of 97%. Li Ronggang et al. [8] used Support Vector Machine (SVM) in the field of crime prediction, comprehensively using logistic regression and SVM to predict the characteristics of suspects. Huang Na et al. [9] based on the improved LSTM network to predict the trend of crime, the average value of the prediction result's RMSE is 127.96, which is 33% higher than the improved LSTM prediction model. Yu Hongzhi et al. [10] used the improved fuzzy BP (Back Propagation) neural network for crime prediction, providing reference for establishing a prediction system with fuzzy concepts.

According to the above review, it can be seen that the number of machine learning model algorithms is increasing. Based on this, this paper will comprehensively use the modeling methods of SVM, Random Forest, XGBoost, and GBDT, and combine them with binary classification to judge the probability of crime occurrence in the Chicago area. Before the experiment starts, we have already counted some data, and the Random Forest (Random Forest) has shown high precision and high recall rate in predicting crime risks. According to a study, the prediction accuracy of the random forest has reached 0.7886, higher than SVM's 0.754 and the neural network's 0.7366. This indicates that the random forest can more accurately identify criminal tendencies compared to other models when predicting crime risks. At the same time, the prediction recall rate of the random forest is also the highest, reaching 0.7192, with higher coverage. Support Vector Machine (SVM) also shows good performance in crime prediction. Although its accuracy is slightly lower than the random forest in some studies, its prediction accuracy has also reached 0.754 [11]. The advantage of SVM lies in its ability to find the optimal hyperplane in high-dimensional space, making it excellent in dealing with complex classification problems. XGBoost (Extreme Gradient Boosting), as an optimization algorithm based on gradient boosting, also shows potential in crime prediction. It constructs multiple weak prediction models and combines them to improve prediction performance. The advantage of XGBoost lies in its ability to handle large-scale datasets and support parallel processing. GBDT (Gradient Boosting Decision Tree) reduces the residuals of the previous model by adding decision trees step by step, improving the prediction performance of the model. It is widely used for classifying samples in the dataset, especially in dealing with non-linear and high-dimensional data.

3. Data Processing and method

The technical route for building and performance evaluation of machine learning-based crime prediction models is shown in Figure 1, consisting of three steps: data preprocessing, model training, model prediction, and performance evaluation.

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

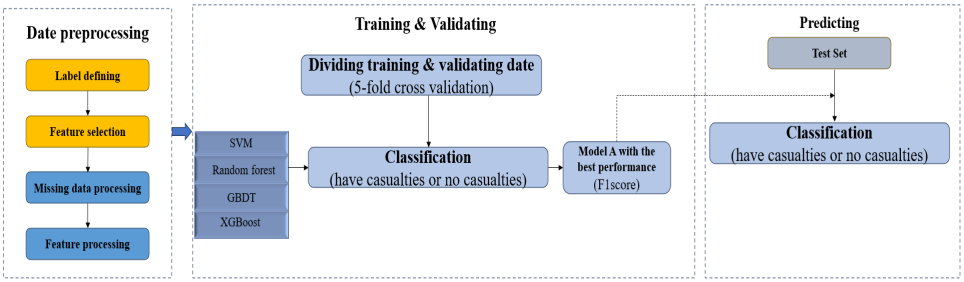

Fig.1. Machine Learning Model Construction and Prediction Performance Evaluation Process.

3.1 Data Processing

The data selected for this study includes 40,000 open-source data of burglary, battery, assault, and criminal damage in Chicago from 2015 to 2020 (data source: https://data.cityofchicago.org/). According to Figure 1, data preprocessing includes four main steps: (1) label definition; (2) feature selection; (3) missing data processing; (4) feature processing. Then, to train and validate the two-step model, the data is first divided into a training set (used to train and validate the model) and a test set (used for prediction). For classification, this study mainly uses binary classification, defining the label of crime occurrence or not, that is, represented by "1" and "0". The entire dataset designated for training serves to develop and assess the efficacy of various machine learning algorithms, including the Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost. A 5-fold cross-validation technique is employed in this process. If there is a numerical value missing one day, the missing value is filled with the average of the previous and next day. Then, the accuracy, recall rate, root mean square error, and F1 score are calculated to select the best-performing model.

3.1.1 Label Selection

Label selection is an important step in machine learning, involving how to assign a label or response variable to each sample in the dataset. This label is the target value we are trying to predict through the model. In classification problems, the label is usually a discrete category, while in regression problems, the label is a continuous value. In this classification step, "crime" (i.e., whether there was a crime) is defined as the label for machine learning. If the event results in any death or injury, the value of this label will be set to "1", otherwise, it will be set to "0".

3.1.2 Feature Selection

Feature selection is the process of selecting the most relevant and informative subset of features from the original feature set. This can help improve model performance, reduce overfitting, and reduce computational costs. The features involved in this study include "District", "DAY", "holiday_or_not", "wind_velocity", "humidity", "air_pressure", "temperature", "mileage_per_day", "number_of_violations_per_day", "affective_index_per_day", "Arrest", "total_rides", "weekend", "AT", "DI", "AT_DI".

3.1.3 Missing Data Processing

In order to improve the overall quality of the data set, ensure the integrity of the data, and ensure that the accuracy of the results will not be affected by missing information during data analysis and modeling, the missing value processing of the data is carried out.To facilitate unified data, this study will take the approach of filling missing values with the average of the previous and next day if there are numerical values missing.

3.1.4 Feature Processing

Feature processing involves transforming and encoding features to meet the requirements of machine learning algorithms. Common methods for handling categorical problems include label encoding (Label Encoding: Convert categorical labels into integers, suitable for ordered categorical data.)

Features used in the classification step must be guaranteed to be numeric. The Label Encoding method is adopted to convert other forms of data into numeric vectors. In this study, the data types of "weekend" and "day" are converted. For example, in this study, weekend is a binary feature, usually with only two values: 0 (non-weekend) and 1 (weekend). Using label encoding, it is directly retained as an integer form: 0 -> 0; 1 -> 1.

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

Day is a categorical feature, indicating a day of the week, usually with seven possible values (e.g., 0=Monday, 1=Tuesday, 2=Wednesday, 3=Thursday, 4=Friday, 5=Saturday, 6=Sunday). Using label encoding, it is retained as an integer form, as shown in Table 1:

Table 1. Label encoding

Day(label encoding)	Day of week
0	Monday
1	Tuesday
2	Wednesday
3	Thursday
4	Friday
5	Saturday
6	Sunday

3.2 Prediction Methods (Classifications)

As shown in Figure 1, we applied four classification methods, namely SVM, RF, GBDT, and XGBoost, and the following describes the application of the four classification methods:

The Support Vector Machine (SVM) is an instructional learning method, predominantly employed for categorization tasks. Its objective is to identify a hyperplane within the feature space that maximally distinguishes data points belonging to different classes. With appropriate adjustments, SVM can be adapted for addressing regression issues, a variant known as Support Vector Regression (SVR). If the gap between any input sample and the hyperplane is at least 1, the hyperplane in n-dimensional space can be characterized as:

$$w^T x + b = 0 \tag{1}$$

where w and x are two different parameters in the hyperplane, w represents the normal vector in n-dimensional space, x represents a point in the hyperplane, and b represents the displacement term or intercept. However, when facing non-linear problems and some linearly inseparable problems, we still need to map them with a kernel function, transforming non-linear problems into linear problems to be solved. Non-linear functions can map non-linear separable problems from the original feature space to a higher-dimensional Hilbert space (Hilbert space) [12], thus transforming them into linear separable problems. At this time, the hyperplane as the decision boundary is represented as follows:

$$w^T \phi(x) + b = 0 \tag{2}$$

where the mapping function is added in the formula. Since the mapping function has a complex form and its inner product is difficult to calculate, the kernel method (kernel method) can be used, that is, the inner product of the mapping function is defined as the kernel function (kernel function) [13].

$$\kappa(X1, X2) = \phi(X1)^T \phi(X2) \tag{3}$$

Random Forest (RF) [14] is an ensemble learning method that improves prediction accuracy and controls overfitting by building multiple decision trees and outputting average results, and it is widely used to solve classification problems. There is no specific formula for Random Forest, but its construction process involves the following steps: (1) Randomly select n samples from the original dataset (replaceable). (2) For each decision tree, randomly select k features for splitting. (3) Build the decision tree until the stopping condition is reached (such as maximum depth or minimum number of samples in the node). (4) For classification problems, output the majority voting result; for regression problems, output the average predicted value.

Gradient Boosting Decision Tree (GBDT) is a collective learning technique that is extensively utilized for categorizing instances within a dataset. It operates by incrementally introducing decision trees to reduce the loss function, with each subsequent tree endeavoring to rectify the mistakes of its predecessor. The fundamental equation for GBDT is:

$$F_m(x) = \sum_{i=1}^{m-1} h_i(x) + h_m(x) \tag{4}$$

where Fm(x) is the prediction result of the m-th tree, hi(x) is the prediction result of the i-th tree, and m is the number of trees.

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

XGBoost is an optimized distributed gradient boosting library designed for speed and performance. XGBoost is an implementation of the Gradient Boosting framework, which minimizes the specified loss function by iteratively training decision trees. XGBoost can be used for binary and multi-class problems. XGBoost mainly includes the following three core formulas:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

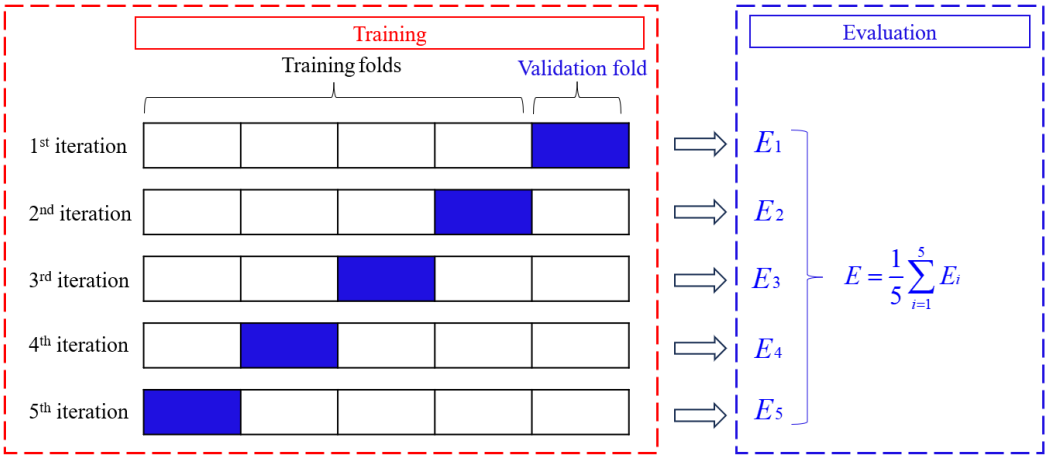

Fig.2. Five-fold cross-validation.

where l is the loss function, y_i is the true value of the i-th sample, $\hat{y}_i$ is the predicted value, K is the total number of trees, f_k is the prediction function of the k-th tree, Ω is the regularization term, and the regularization term can be explained as follows:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{6}$$

where, γ is the complexity parameter of the tree, T is the number of leaf nodes of the tree, λ is the L2 regularization parameter of the leaf node weights, w_j^2 is the weight of the j-th leaf node. The prediction function can be explained as follows:

$$f(x) = \sum_{k=1}^{K} w_k I(x \in D_k) \tag{7}$$

where, w_k is the weight of the k-th tree, $I(x \in D_k)$ is the indicator function, if x belongs to the j-th leaf node of the k-th tree, then the value is 1, otherwise it is 0.

3.2.1 Five-fold cross-validation

Five-fold cross-validation is a model evaluation method, a special case of k-fold cross-validation where k equals 5[15]. This method divides the dataset into 5 equal parts (or as equal as possible), and in each iteration, one part is used as the test set while the remaining four parts are combined as the training set, as shown in Figure 2. For example, in iteration 1: the first part is used as the test set, and the second, third, fourth, and fifth parts are used as the training set. In each iteration, the model is trained on the training set and its performance is evaluated on the test set. The average value of the model performance metrics across all iterations is then calculated to obtain the final evaluation result of the model. Five-fold cross-validation is a commonly used cross-validation technique, but in some cases, other numbers of folds (such as ten-fold cross-validation) can be chosen to more comprehensively assess model performance.

3.2.2 Classification Metrics: Definitions and Explanations

Precision is a metric that measures the ability of a machine learning model to predict positive samples accurately [16]. This metric reflects the ratio of correctly identified true positives to the total number of instances classified as positive by the model. An elevated precision value signifies superior predictive precision, suggesting that the model is more successful in correctly classifying positive samples when it

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

predicts them as such.

Recall, also known as hit rate, is a metric used in machine learning to measure the performance of a model, particularly in classification problems. It refers to the proportion of actual positive samples that the model successfully predicts as positive.

The F1-Score is a statistical measure used to evaluate the precision of binary classification models [17]. It combines the precision and recall of a classification model and is their weighted harmonic mean. The F1-score is applicable to binary classification problems, especially useful in cases where the positive and negative samples are imbalanced, providing a more comprehensive assessment of model performance.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{10}$$

where TP (True Positive) represents the number of the correctly classified positive samples, which means the number of crime cases that occurred and were correctly predicted by the machine learning model. Similarly, FP (False Positive) represents the number of the incorrectly classified negative samples, which means the number of crime cases that did not occur but were incorrectly predicted as occurring. FN (True Negative) represents the number of the incorrectly classified positive samples, which means the number of crime cases that occurred but were incorrectly predicted as not occurring.

4. Crime prediction results

In the Chicago crime forecasting model, the outcomes of the performance assessment during the classification phase are presented in Table 2. The metrics of precision, recall, and F1-score are determined for both the training dataset and the test dataset. The figures in this table represent the mean of the outcomes when considering both negative instances (indicating the presence of crime) and positive instances (indicating the absence of crime). From the data listed in the table, the numbers outside brackets in Table 2 represent the scores of the assessment indicators on the test set and the numbers in brackets represent the scores on the training set. From the above experimental results, it can be seen that GBDT and XGBoost are superior in the classification model of machine learning for predicting whether crime occurs. We can clearly see that XGBoost (0.69) and GBDT (0.71) have the highest precision in the F1-score. The test set and the training set are also very close on the other two data items (Precision and Recall), which means the model has strong generalization ability and can adapt well to different data distribution. Therefore, in the future machine learning, we recommend GBDT and XGBoost. The predictive performance of each model is as follows:

(1) GBDT model has high prediction accuracy and generalization ability. In the experiment, the input data is trained and the model is evaluated with the data of the test set. In contrast, other models have a large average absolute error, which can not accurately reflect the actual situation. GBDT model also has good generalization ability and stable prediction performance on different data sets.

(2) GBDT model has strong integrated learning ability. GBDT uses a gradient lifting algorithm, which can integrate multiple weak classifiers to form a strong classifier, thereby improving prediction performance. At the same time, the weak learner generated by the previous learner will also have an impact on the next learner, strengthening the learning where the previous weak learner predicted errors. Compared with GBDT, random forest and SVM models may be limited in dealing with complex relationships.

(3) GBDT model features selection and model interpretation ability is strong. GBDT can sort the features according to their importance, select the features with better classification effect, and assign corresponding weights to the features, so as to improve the classification effect of the model. GBDT can also calculate how much each feature contributes to classification predictions, helping people better understand the model's decision-making process.

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

In general, according to the experimental results, it can be concluded that the GBDT model is better than XGBoost, random forest and SVM models in predicting the occurrence of crime cases. The ability of nonlinear modeling, the ability of capturing complex data relationships, the ability of ensemble learning, the ability of generalization and the ability of feature selection make GBDT an ideal classification prediction model.

Table 2. Performance of the classification models

Model	Precision	Recall	F1-score
SVM	0.79(0.81)	0.67(0.72)	0.63(0.70)
RF	0.73(0.79)	0.57(0.66)	0.48(0.62)
XGBoost	0.70(0.69)	0.69(0.68)	0.69(0.68)
GBDT	0.71(0.71)	0.71(0.71)	0.71(0.71)

5. Conclusions

In this paper, four machine learning algorithms are applied to predict crime in Chicago—Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Decision Tree (GBDT), and XGBoost—to predict the probability of crime occurrence in Chicago from 2015 to 2020, focusing on crimes such as burglary, battery, assault, and criminal damage. Through five-fold cross-validation and a variety of performance evaluation metrics (precision, recall, and F1 score), it was found that these four models all demonstrated high accuracy and recall rates in predicting crime risks. Although there were slight performance differences between the training and testing sets for each model, the F1 scores were very close overall, indicating that these models possess high reliability and effectiveness in predicting the occurrence of crimes.

In addition, the advantages of machine learning binary classification methods in the field of crime prediction are mainly reflected in the following aspects: Firstly, machine learning algorithms can handle a large amount of historical crime data, socio-economic indicators, and geographical information, and establish models to predict the timing, location, and type of crime, thus achieving a shift from passive response to proactive prevention. Secondly, machine learning models, especially Random Forest and Support Vector Machine, have shown high precision and recall rates in predicting crime risks, effectively identifying criminal tendencies. In addition, algorithms such as XGBoost and GBDT have advantages in processing large-scale datasets and supporting parallel processing, which can enhance predictive performance.

However, this experiment only involved the classification step of machine learning and did not involve more complex algorithms like regression, which may affect the scientific nature of the results to some extent. Moreover, the generalization ability of the models is limited by the quality and quantity of the training data [18]; if the training data is biased or incomplete, the predictive results of the models will also be affected. Machine learning models are often seen as “black boxes,” lacking transparency in their decision-making processes [19], which can be an issue in police work that requires explanation and trust. Furthermore, as the volume of data increases and the complexity of the models rises, the computational costs and time will correspondingly increase, which may limit the application of the models in real-time crime prediction. Lastly, machine learning models require regular updates and maintenance to adapt to changes in crime patterns, necessitating ongoing resource investment and professional knowledge to address the evolving policing paradigms of the future.

It is true that the opacity of machine learning algorithms, the complexity of data-driven decisions, the hidden dangers of intellectual property and trade secrets involved, and the challenges of predictive policing seem to make machine learning algorithms more of a crisis of trust in police work, but it is precisely this that drives the development of interpretable artificial intelligence (XAI). Machine learning we can temporarily think of as its predecessor, when the development of interpretable artificial intelligence (XAI) to a certain extent, we worry about the problem may not exist today, it is worth our future to do further research.

References

[1] HE Rixing, LU Yumei, JIANG Chao, DENG Yue, LI Xinran, SHI Dong. Progress in Research and Practice of Spatial-temporal Crime Prediction over the Past Decade[J]. Geo-Information Science, 2023, 25(4): 866-882

ISSN 2522-3488 Vol. 9, Issue 1: 7-14, DOI: 10.25236/IJNDES.2025.090102

[2] Yan Jinghua, Hou Miaomiao. Research on Time Series Prediction of Theft Crime Based on LSTM Network [J]. Data Analysis and Knowledge Discovery, 2020, 4(11): 84-91.
[3] Shen Hanlei, Zhang Hu, Zhang Yaofeng, et al. Research on Burglary Crime Prediction Based on Long Short-Term Memory Model[J]. Statistics and Information Forum, 2019, 34(11): 107-115.
[4] Han Yishi, Fan Yingsheng, Li Guojun, et al. Analysis of the Growth Trend of Communication Network Fraud Crime Based on ARIMA Model: A Case Study of Quzhou City, Zhejiang Province[J]. Theoretical Observation, 2017, (05): 101-103.
[5] Hou Miaomiao, Hu Xiaofeng. Crime Prediction Research Based on Time Series Model SARIMA[J]. Journal of People's Public Security University of China (Natural Science Edition), 2021, 27(02): 67-73.
[6] Liu Xiaojuan, Gao Liansheng. Application of Grey System Theory in Dynamic Crime Prediction[J]. Journal of People's Public Security University of China (Social Science Edition), 2005, 66(1): 44-48.
[7] Alves L G, Ribeiro H V, Rodrigues F A. Crime Prediction Through Urban Metrics and Statistical Learning [J]. Physica A: Statistical Mechanics and Its Applications, 2018, 505: 435-443.
[8] Uttley J, Canwell R, Smith J,et al.Does darkness increase the risk of certain types of crime? A registered report protocol[J].PLoS ONE (v.1;2006), 2024, 19(1):18. DOI:10.1371/journal. pone. 0291971.
[9] Huang Na, He Jingsha, Sun Jingchao, et al. Crime Trend Prediction Method Based on Improved LSTM Network[J]. Journal of Beijing University of Technology, 2019, 45(08): 742-748.
[10] Yu Hongzhi, Liu Fengxin, Zou Kaiqi. Improved Fuzzy BP Neural Network and Its Application in Crime Prediction[J]. Journal of Liaoning University of Engineering and Technology (Natural Science Edition), 2012, 31(02): 244-247.
[11] Mwaniki B, Mwalili T, Ogada K. Crime Prediction Using Decision Trees, Random Forests, and Hybrid Algorithm: A Comparative Analysis[C]//Proceedings of the 2023 7th International Conference on New Media Studies (CONMEDIA), Bali, Indonesia, 2023: 99-104.
[12] Zhou Z H. Machine Learning [M]. Beijing: Tsinghua University Press, 2016: 121-139, 298-300.
[13] Li H. Statistical Learning Methods[M]. Beijing: Tsinghua University Press, 2012: Chapter 7, 95-135.
[14] Biau G, Devroye L. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification[J]. Journal of Multivariate Analysis, 2010, 101: 2499-2518.
[15] Kim Y A, Wo J C, Hipp J R .Estimating Age-Graded Effects of Businesses on Crime in Place [J]. Justice Quarterly, 2023, 40(3):403-426. DOI:10.1080/07418825.2022.2107943.
[16] Rainio O, Teuho J, Klén, Riku. Evaluation metrics and statistical tests for machine learning[J]. Scientific Reports, 14(1):1-14[2024-11-26].
[17] Hand D J, Christen P, Kirielle N .F: an interpretable transformation of the F-measure[J]. Machine Learning, 2021(1).*
[18] Mandavkar Y K .Crime Analysis and Prediction using Data Mining[J].International Journal of Advanced Research in Science, Communication and Technology, 2023.DOI:10.48175/ijarsct-12027.
[19] Parisineni S R A, Pal M .Enhancing trust and interpretability of complex machine learning models using local interpretable model agnostic shap explanations[J].International Journal of Data Science and Analytics, 2024, 18(4):457-466.