Research Article

A Prediction Model of Recidivism of Specific Populations Based on Big Data

Jing Leng,1,2 Wei Xu,1,2 Tonghong Li,1,2 Likun Chen,1 and Mingwei Xu3

1Department of Information Technology, Hubei University of Police, Wuhan, 430000 Hubei, China
2Electronic Forensics and Trusted Applications Hubei Collaborative Innovation Center, Wuhan, 430000 Hubei, China
3School of Computer Science and Technology, Tiangong University, Tianjin 300000, China

Correspondence should be addressed to Wei Xu; weixu7810@163.com and Tonghong Li; lthong@hbpa.edu.cn

Received 18 April 2022; Revised 6 May 2022; Accepted 11 May 2022; Published 25 May 2022

Academic Editor: Kuruva Lakshmanna

Copyright © 2022 Jing Leng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This research is aimed at establishing and improving the relevant information database of “released prisoners,” integrating the behavior data of specific populations, constructing the early-warning model of recidivism, evaluating the database file information by using prediction analysis technology, and making a prediction alarm to the people who are most likely to commit the crime again on the basis of big data analysis so as to ultimately achieve the goal of reducing the recidivism rate. The research used the data exchange technology of the heterogeneous database to complete data collection and database establishment, used the feature engineering technology to analyze the big data of specific populations, obtained the multidimensional behavior trajectory data, and carried out sorting and statistics. On this basis, the linear regression algorithm was applied to make the prediction and evaluation, and the visual results were presented to assist in researching and judging the possibility of recidivism of specific personnel. Through programming realization and simulation experiments, the research obtained the tendency prediction of the people to commit crimes again by statistical analysis from multidimensional data with a long time span. In the next step, the real test will be carried out to help the public security work in China and contribute to the maintenance of national and social stability.

1. Introduction

The resettlement, assistance, and education of released prisoners is a social project that has always been attached importance by the Central Committee of the CPC (Communist Party of China) and the State Council, which is of great significance to maintain social stability, reduce recidivism, and ensure and promote the smooth progress of reform and opening up. In recent years, various regions and relevant departments have done a great deal of fruitful work in this regard, making positive contributions to economic development and social stability. However, from the perspective of practical work, the current work does not adapt to the development of the situation. The main reason is that the basic information work on the released prisoners and the tracking of the later situation are relatively weak, and the resultant force of all aspects of society has not been really formed.

The number and rate of recidivism of released prisoners are increasing year by year. According to the statistics of prisoners in six prisons from 2010 to 2015 in a province, there were 2761 criminals who had been “put into prison for the second time,” accounting for 5.18% of the prisoners. In the past five years, 458, 434, 501, 617, and 751 prisoners were sentenced to fixed-term imprisonment for committing crimes again, with the recidivism rates of 4.57%, 4.36%, 4.63%, 5.65%, and 6.46%, respectively. 91.27% went to prison for the second time, 7.7% for the third time, 0.80% for the fourth time, and 0.14% for the fifth time.

The released prisoners with poor performance in prison are more likely to commit crimes again. While in prison, the original criminal psychology of these released prisoners has not been well corrected, their criminal thoughts have not been thoroughly reformed, and their bad habits have not been fundamentally corrected. They do not want to repent, adhere to the criminal stand, attribute the punishment to...
social injustice, and attribute the responsibility to the prose-
cutors, victims, and other social members. They have a
strong antisocial consciousness and a serious psychology of
revenge against society. After they get out of prison, they will
continue to commit crimes and endanger society.

It is a complicated social problem for the prisoners to
commit crimes again. With the continuous development of
the social economy, the recidivism of released prisoners
has gradually become an important factor affecting social
security and stability. The rate of recidivism has become an
important indicator reflecting the social security and stabil-
ity of a country or region and the effect of prison correction
work. Compared with the developed western countries, the
recidivism rate of released prisoners in China is relatively
low, but the current situation is still not optimistic. Espe-
cially in the period of comprehensive and rapid development
in China, the characteristics, situation, and great social harm
of the recidivism of released prisoners must be paid enough
attention to, which is of great practical significance to the
stability of social order, the protection of the country and
people’s lives and property, and the promotion of social
development [1].

The recidivism of released prisoners seriously disturbs
the social order, infringes on the security of the country
and the safety of the life and property of the people, and
destroyes the social and economic development, which has
become an essential factor affecting the long-term stability
of the country. To take positive and effective preventive
measures to reduce the recidivism of released prisoners
and promote the healthy, orderly, and stable development
of society is not only a key aspect of education and reform
in prisons but also the social responsibility of governments
at all levels, relevant departments and family relatives of
released prisoners.

The countermeasures to prevent the released prisoners
from committing crimes again are as follows. (1) The con-
struction of the legalization, scientization, and socialization
of the prison work should be actively promoted, the reform
of the prison system should be well done, the education and
reform of prisoners out of prison should be effectively
strengthened, and the quality of the education and reform
of prisoners should be strengthened. (2) The social manage-
ment, resettlement, assistance, and education of released
prisoners should be strengthened. (3) Family quality and
family relations should be improved to create a good family
environment.

As for the second point above, the research team of
this project has conducted investigations and studies in
prison many times and found that there is a lack of a uni-
fied information management system platform connecting
various social departments for such personnel in actual
work, and there is no prediction and early warning for
their recidivism [2].

Therefore, it is necessary to integrate the information of
all sectors of society to strengthen the supervision, assis-
tance, and education of the released prisoners, timely link
up and feed back the work of relevant departments, and
implement the resettlement, assistance, and education work
of relevant departments and grassroots streets and commu-
nities. To reduce the social crime rate and help the public
security organs safeguard the safety of the country and the
people, the research and application of the prediction and
early-warning model for the recidivism of the released pris-
oners are of great significance [3].

At present, western countries begin to use advanced
“predictive analysis” technology to monitor specific areas
and populations and carry out crime prediction and preven-
tion. "Blue CRUSH,” a predictive software program used by
the police department of Memphis, Tennessee, uses histori-
cal data to reduce crimes. It works by analyzing crime and
arrest data and combining it with weather forecasts, eco-


nomic indicators, and information about events such as pay-
days and concerts. Since the police department began using
the software in 2005, the crime rate in Memphis has dropped
by 30%. The U.S. Department of Justice began to use predic-
tive analysis technology to evaluate the data in its crime
assessment system to help predict which released prisoners
are most likely to commit crimes again. In Florida, the U.S.
Department of Justice recently began using the same soft-
ware to help predict which juvenile offenders are likely to
continue to be repeat offenders, marking these as specific
prevention and education programs. The patent of this
product belongs to IBM (International Business Machines)
company in the United States, and some cities in the UK
(United Kingdom) are experimenting with this product [4].
At present, some related research institutions are developing
this kind of technology, but no similar results have been pro-
duced [5].

2. Relevant Work

2.1. Demand Analysis. The cohesion and cooperation of
the relevant departments should be strengthened to ensure the
implementation of resettlement, assistance, and education
work and comprehensively improve the overall level of the
resettlement, assistance, and education work for released
prisoners. In addition, it is necessary to strengthen the clas-
sified management of the released prisoners so that a small
number of these people who may be harmful to society are
always under the supervision of society. These people should
also be included in the management of resettlement, assis-
tance, and education work, and the relevant departments
should strengthen the dynamic management measures to
prevent management negligence.

This system will connect prisons, judicial administration
departments, public security organs, human resources and
social security departments, industrial and commercial
administration departments, civil affairs departments, grass-
roots streets, and communities [6].

The judicial administration department cooperates with
all relevant units. On the one hand, it extends the work of
resettlement, assistance, and education to the prison; on
the other hand, it also closely connects with the social reset-
tlement, assistance, and education work and gradually insti-
tutionalizes and procedures. All relevant functional
departments shall perform their own duties and actively do
a good job in the specific connection and cooperation of
resettlement, assistance, and education work.
Half a month before the released prisoners leave the prison, the prison should hand over their performance and other relevant materials to the local public security organ where they are registered and strictly perform the handover procedures. A system of regular inspection and local feedback should be established for those released from prison within three years.

The labor department should actively assist the grassroots party and government organizations in the streets and townships to carry out employment guidance and skill training for those who have not yet been released from prison and do a good job in various employment services to create conditions for their employment. For the on-the-job workers who have participated in the unemployment insurance, when they are sentenced and are unemployed after they are released from prison, the unemployment relief fund shall be issued according to the provisions of the unemployment insurance for workers. For those who are sentenced during the period of enjoying unemployment relief, after they are released from prison, they may continue to receive unemployment relief funds according to the period for which they have not finished enjoying the relief.

When the released prisoners who are not yet employed apply in accordance with the law to engage in self-employed industrial and commercial operations or to establish other economic entities, the industrial and commercial administration department should give equal treatment to them and protect their legitimate rights and interests. The department should also strengthen education and management through the relevant organizations, implement the public security responsibility system, enhance their concept of law-abiding management, and improve the level of professional ethics.

The civil affairs department should encourage the economic entities set up by the townships, streets, and village committees to place the released prisoners and mobilize the grassroots organizations of political power to actively participate in and do well in the work of resettlement, assistance, and education. Through the form of community service, the work of helping and educating the delinquent youth is included in the community service series.

The public security organs should establish a responsibility system of helping and educating the released prisoners and actively cooperate with the prison to do a good job in the regular assessment of the quality of the personnel transformation. Especially for those who have bad habits and poor effect of transformation and have the tendency to commit crimes again, the public security organs should strengthen management, closely grasp their ideological trends and action directions, and strive to avoid management negligence. For those who commit crimes again, they should be punished seriously according to the law.

2.2. The Construction of the Platform and Information Exchange Center. Using the advanced information management technology of the Internet, the core business data center, information exchange center, and unified information management platform for specific personnel management are established to effectively collect and feed back the relevant information, transformation performance, resettlement, assistance, and education data of the released prisoners. The information data of each administrative unit and the information of other business places should be integrated, and the tracking, grading, and classified management of key personnel should be strengthened. In addition, it is necessary to clarify the division of labor among departments and strengthen the cohesion and cooperation in the work of relevant departments so as to ensure the implementation of resettlement, assistance, and education.

It provides a data support platform for the research of “how to take positive and effective preventive measures to reduce the occurrence of recidivism cases of released prisoners” [7].

The overall structure design of the information management platform for specific populations is shown in Figure 1. The platform applies data exchange technology based on the cross-department heterogeneous database to complete data collection and database establishment and filters, transforms, and classifies the collected information [6]. Based on the platform, the prediction and analysis system of recidivism is used to analyze and evaluate the information data of the released prisoners. Compared with the historical data of recidivism, the predictable model of recidivism is established so as to make the prediction and early warning for the released prisoners who are most likely to commit the crime again and effectively manage and monitor them so as to reduce recidivism [8, 9].

3. The Construction of the Crime Prediction and Analysis Model

According to the data platform, the model of crime prediction and analysis is established to reduce recidivism [10, 11].

(1) One Theme. Standardized, informationized, networked, and scientific management should be carried out on the settlement, assistance, and education of those released from prisons, and scientific and effective decision-making should be provided to prevent the occurrence of such people’s recidivism.

(2) Establish Two Centers. A core business data center and an information exchange center should be set up for the management of released prisoners.

(3) Realize Five Functions. Realize the report analysis and auxiliary decision-making function, personnel information management function, major matters management function, unified information portal and application integration function, and information security protection function.

3.1. Heterogeneous Data Standard. The data required by the platform comes from prisons, judicial administration departments, public security organs, human resources and social security departments, industrial and commercial administration departments, civil affairs departments, grassroots streets, community-related personnel, transportation, banks, consumption, and so on. The amount of information is huge and diverse. To collect and organize massive business data, Logstash of Elastic Stack technology is used to collect data and establish a unified data entry.
In Figure 2, “input” defines how to collect data, “filter” defines data processing, and “output” defines data storage.

3.2. Detailed Design of the Prediction Model. The overall idea of building the prediction model for the recidivism of the released prisoners is shown in Figure 3. Through the population information, water, electricity, and gas information, medical treatment information, hotel information, travel information, bank account statement information, payment information, work information, communication data, and network data of the released prisoners, the prediction analysis model is built by using machine learning to ensure the accuracy of the recidivism prediction and provide a certain reference for reducing recidivism [12, 13].

3.2.1. Data Preparation. Based on the specific population information data of prisons, judicial administration departments, public security organs, human resources and social security departments, industrial and commercial administration departments, civil affairs departments, grassroots streets, communities, and other departments, features are extracted from multiple dimensions for modeling [14].

Data processing is the operation of cleaning and deduplication of the original data. Data cleaning includes invalid data filtering, missing value filling, and data conversion. Data comes from different databases, and the data standard is not consistent, so data needs to be standardized [15].

3.2.2. Feature Engineering. Feature engineering is a process that uses professional background knowledge and skills to process data, which makes features play a better role in machine learning algorithms. Data and features determine the upper limit of machine learning, while the algorithm and model are only close to this limit. It can transform any data (such as text or image) into digital features that can be used for machine learning.

If the collected data is in the form of text, the feature extraction should be used to transform the text into data. However, the mathematical formula in the machine learning algorithm cannot effectively identify the text, writing, and string, so the text type should be converted into a numerical value.

The dimensionless method is to transform the original data to the range of the mean of 0 and standard deviation of 1 by means of data standardization. In feature dimensionality reduction, the uncorrelated main variables are deleted and low variance filtering is performed.

Feature statistics is to analyze and compare the behavior data of specific personnel with that of ordinary people and to make feature statistics on the behavior data of specific personnel. Then, the quality of eigenvalues is enhanced by using feature cleaning, feature transformation, and other technical methods to provide eigenvalues for the prediction model [16, 17].

3.2.3. Data Segmentation. To verify the effect of the model, the best model is selected, and the dataset is divided into a training set and a test set. The training set is used to train the model, and the test set is used to test the effect of the model. This dataset is segmented by common segmentation proportions and methods; that is, by random 7:3 segmentation, 70% is used for training and 30% for testing [18].

3.2.4. Model and Evaluation. The process of modeling is a process of continuous testing and exploration [19]. The population information, water, electricity, and gas information, medical treatment information, hotel information, travel information, bank account statement information, payment information, work information, communication data, network data, and other eigenvalues are used to establish the model. To verify the effect of the model, the best model is selected, and the dataset is divided into a training set and a test set. The training set is used to train the model, and the test set is used to test the effect of the model [20].

There are two methods for model evaluation. The first method is the real value and the predicted value. The training set and the test set have been divided before, and the predicted values are used to compare whether the real results in the test set are consistent with the target values. The second method is the calculation accuracy. The evaluation index is calculated based on the confusion matrix. The true-positive sample is the positive sample that is correctly classified by the model; the false-negative sample is the positive sample that is wrongly classified by the model; the false-positive sample is the negative sample that is wrongly classified by the model; the true-negative sample is the negative sample that is correctly classified by the model. The evaluation index is based on the confusion matrix, as shown in Table 1 [21, 22].
The equation of accuracy is as follows:

\[
\text{Accuracy rate} = \frac{TP + TN}{TP + TN + FP + FN}.
\]  

4. Total Prediction Process

The platform first collects multisource heterogeneous data, then cleans data, filters invalid data, fills missing values, and converts data, which can reduce unnecessary interference analysis of dimensions. Data deduplication is to avoid leaving duplicated data after cleaning [23].

The feature engineering module extracts the behavior data of specific personnel. The basic quality and health indicators of family members, the cost of water, electricity, and gas per month, the medical treatment cost and physical condition indicators, the bank account statement, cash, the expenditure expenses of WeChat and Alipay, the frequency of work, the number of calls per month, and the number of times of Internet access are used for analysis. The data is divided randomly by 7:3, which is 70% for training and 30% for testing.

The feature extraction in feature engineering is used to transform any data (such as text or image) into digital features that can be used for machine learning. Then, dimensionless processing is carried out for the features with a large difference in unit or size. On this basis, feature dimensionality reduction and feature statistics are carried out. Finally, the linear regression algorithm is used for prediction and evaluation [24]. The overall process is shown in Figure 4.

5. Core Methods

5.1. Linear Regression Algorithm. The core algorithm of this research is the linear regression algorithm. Linear regression is a linear model to predict through the linear combination of attributes, mainly to find a straight line or a plane or a higher dimensional hyperplane so as to minimize the error between the predicted value and the real value.
Linearity means that the relationship between two variables is a linear function, and the image is a straight line. Regression means that when people measure things, due to the limitation of objective conditions, they always get the measured value rather than the real value of things. In order to get the real value, they carry out infinite measurements and finally calculate and return to the real value through these measured data. The problem solved by linear regression is to process a large number of observation data so as to get the mathematical expression, which is more consistent with the internal law of things, and to find the law between data so as to simulate the results, which is to predict the results. The solution is to get unknown results from known data. Advantages are as follows: the results have good interpretability, and the calculation of entropy is not complex. Disadvantages are as follows: the nonlinear data fitting is not good. The applicable data types are numerical data and nominal data.

In the linear regression model, there are univariate linear regression, multivariate linear regression, and the extended generalized linear model. The univariate linear regression is the simplest form of regression. A variable represented in the two-dimensional plane coordinate system is a pile of discrete points, and the best straight line is found to fit these points (as shown in Figure 5). These known discrete points are the training dataset. Thus, for a new unknown piece of data, when only its abscissa is known, its ordinate can be predicted according to the model.

Many variables are processed, and the results are used as independent variables for prediction analysis [25], as shown in the following equation:

\[ Q_t = a q_t + b. \]  

\[ (2) \]

The independent variable value of the \( t \)-th period is expressed by \( q_t \), \( Q_t \) represents the dependent variable value of the \( t \)-th period, \( a \) can be obtained by equation (3), and \( b \) can be obtained by equation (4):

\[ a = \frac{\sum q_t}{n} - b \frac{\sum q_t}{n}, \]

\[ (3) \]

\[ b = \frac{n \sum q_t Q_t - \sum q_t \sum Q_t}{n \sum q_t^2 - (\sum q_t)^2}. \]

\[ (4) \]

Input a set of data \( x \) and \( y \) into the model, and the line segment as shown in Figure 6 can be obtained.

The linear fitting does not cross all points, so it is necessary to optimize the model and introduce the loss function to estimate the inconsistency between the predicted value \( f(x) \) and the real value \( Y \). The smaller the loss function is, the better the effect of the model will be. The mathematical equation model of the loss function is shown in the following equation:

\[ f(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2n} \sum_{i=1}^{m} \left( h \theta^{(i)} - y^{(i)} \right)^2. \]

\[ (5) \]

The average value of the sum of squares of (predicted value – real value), i.e., the sum of the distance from the point to the line, is the minimum.
Data processing includes the filling of missing values, data conversion, and data deduplication. The main program libraries used in the development process are NumPy, Pandas, and SciPy. In general, NumPy and Pandas need to be used in combination. First, the prepared original data is read. According to the missing rate and importance of missing values, the processing strategies are shown in Table 2. The data conversion subprocess mainly deals with the format conversion of the data. Some floating-point fields are directly converted to integer fields, which will not have a great impact on the training of the model. Finally, the data need to be reprocessed and the same data need to be deleted.

The original data is processed into semistructured data, but it cannot be used in feature engineering and model training. If all the data are used for model training, a new batch of data will need to be collected for the accuracy test of the model, resulting in greater uncertainty. Therefore, it is necessary to use the Sklearn library to divide the data by a random 7:3 ratio. The Sklearn library is a simple and efficient data mining and analysis tool, which is based on NumPy, SciPy, and Matplotlib.

The feature engineering module mainly includes feature extraction, the dimensionless method, and feature statistics, which is realized by Sklearn. The dimensionless method is to transform the collected data to the range of the mean of 0 and standard deviation of 1 by means of data standardization. The related functions of data standardization processing have been encapsulated in Sklearn, and the function of StandardVectorizer is called. For feature statistics, the released prisoners return to normal life, and the corresponding data records will be generated every day, mainly referring to the water, electricity, and gas information, medical treatment information, hotel information, travel information, payment information, etc., and the relevant data will be counted by month. It mainly uses the CounterVectorizer function in Sklearn.

Through the above steps, the original data is processed into semistructured data, then transformed into structured data through feature engineering, and then input into the model for training, mainly using the linear regression method in the Sklearn library for linear regression prediction. At this time, a large amount of data is needed to train the model, which is time-consuming and has certain requirements for the physical machine, so it can be put in the background. The output results are displayed by using Python’s 2D drawing library, Matplotlib, which generates graphics that can achieve publishing quality in various hard copy formats and cross-platform interactive environments [26, 27].

6. Experimental Simulation

6.1. Experimental Environment. The implementation and testing of this model is mainly based on Python, which supports functional programming and OOP (Object-Oriented Programming) and can develop all kinds of software. Due to the development of NumPy, Matplotlib, Scrapy, PyTorch, and other libraries, Python is more and more suitable for scientific calculation and prediction, including but not limited to machine learning [28], data cleaning and analysis, neural networks, and artificial intelligence. The development environment of this research is shown in Table 3.

The Tkinter library is mainly used to develop a user interface [29], which is very stable and has a small extra cost. It is one of the standard GUI libraries of Python.

6.2. Experimental Process. A simulation experiment was carried out on the platform to analyze and judge the behaviors of person A and person B in the first half of 2020. The multidimensional behavior data of the year were processed and presented in the form of a scatter graph by visualization technology.

The platform was used to simulate and collect the behavior data of A and B in 12 months in 2020, including 14 dimensions of data, such as water, electricity, and gas information, medical treatment information, hotel information, travel information, and bank account statement information, as shown in Tables 4 and 5.

In the tables, data items that can be obtained through public channels, such as “water fee,” “electricity fee,” and “gas fee,” can be queried on the company platform corresponding to each resource. The number of times taken by train and other means of transportation can be replaced by the number of tickets purchased by querying the ID number. The employment of special personnel is usually registered in the neighborhood office where they live. The hotel accommodation is bound with the ID number for the public...
security department to keep. Therefore, these data can be obtained by querying the public department.

The private data items that cannot be obtained through public channels, such as "bank account statement," "phone," "surf the Internet," "cash," "Alipay," and "Webcat," cannot be obtained from the data directly from the relevant company. Only after the person is considered to be suspected of a crime and the public security organ performs the case handling procedures in accordance with the law can he apply to the relevant company for inquiry of these data. Among them, WeChat and Alipay all refer to the financial savings function, so the unit of this item is "yuan."

The collected data was used as the input of the recidivism prediction model. After data processing, the dimensionless method, feature engineering, and other processing,

Table 2: General processing strategy of missing data.

<table>
<thead>
<tr>
<th>Importance</th>
<th>Low</th>
<th>Miss rate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Simple filling or no processing</td>
<td>Remove the field directly</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>Fill through calculation; estimate through experience or business knowledge</td>
<td>Try to complete from other channels; use other fields to get by calculation; remove the field and mark it in the result</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Introduction of the development environment.

<table>
<thead>
<tr>
<th>System/software name</th>
<th>Version</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows 64 bit</td>
<td>1909</td>
<td>This software runs on the Microsoft operating system</td>
</tr>
<tr>
<td>PyCharm Community Edition 2019.1.3</td>
<td>2019.1.3</td>
<td>IDE is mainly used in the process of software implementation</td>
</tr>
<tr>
<td>Python</td>
<td>3.8.1</td>
<td>Interpreter version</td>
</tr>
<tr>
<td>NumPy</td>
<td>1.18.1</td>
<td>Data analysis library</td>
</tr>
<tr>
<td>Pandas</td>
<td>1.0.3</td>
<td>Data analysis library</td>
</tr>
<tr>
<td>Tkinter</td>
<td>8.6.0</td>
<td>User interface library</td>
</tr>
<tr>
<td>Matplotlib</td>
<td>3.2.1</td>
<td>Graphic visualization library</td>
</tr>
<tr>
<td>Jieba</td>
<td>0.42.1</td>
<td>Data analysis library</td>
</tr>
<tr>
<td>Sklearn</td>
<td>0.23.1</td>
<td>Machine learning library</td>
</tr>
<tr>
<td>SciPy</td>
<td>1.4.1</td>
<td>Scientific computing library</td>
</tr>
</tbody>
</table>

Table 4: Behavior data of person A in 2020.

(a)

<table>
<thead>
<tr>
<th>Months</th>
<th>Airplane (times)</th>
<th>High-speed train (times)</th>
<th>Bus (times)</th>
<th>Taxi (times)</th>
<th>Water fee (yuan)</th>
<th>Electricity fee (yuan)</th>
<th>Gas fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>51</td>
<td>55</td>
<td>49</td>
<td>55</td>
</tr>
<tr>
<td>February</td>
<td>51</td>
<td>51</td>
<td>50</td>
<td>52</td>
<td>55</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>March</td>
<td>55</td>
<td>51</td>
<td>53</td>
<td>52</td>
<td>60</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>April</td>
<td>53</td>
<td>55</td>
<td>52</td>
<td>52</td>
<td>66</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>May</td>
<td>57</td>
<td>54</td>
<td>53</td>
<td>55</td>
<td>65</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>June</td>
<td>65</td>
<td>58</td>
<td>54</td>
<td>56</td>
<td>66</td>
<td>66</td>
<td>60</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Months</th>
<th>Population quality (high 2)</th>
<th>Health (good 2)</th>
<th>Medical treatment costs (yuan)</th>
<th>Physical condition (1 is good)</th>
<th>Working condition (normal 2)</th>
<th>Frequency of job change (frequent 2)</th>
<th>Hotel check-in (times)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>2</td>
<td>2</td>
<td>500</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>February</td>
<td>2</td>
<td>2</td>
<td>501</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>March</td>
<td>1</td>
<td>2</td>
<td>502</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>April</td>
<td>1</td>
<td>1</td>
<td>503</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>May</td>
<td>1</td>
<td>1</td>
<td>504</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>June</td>
<td>1</td>
<td>1</td>
<td>505</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
The linear regression algorithm was used for linear regression prediction and model evaluation to obtain the predicted value of the recidivism rate in the first half of 2020. The visualization results are shown in Figure 7. The blue points in the figure are the predicted value of the corresponding month, the yellow line is the trend line of recidivism, the horizontal axis is the month, and the vertical axis is the predicted value. If the predicted value is close to the trend line of recidivism, it proves that the prediction is accurate; if the predicted value fluctuates greatly and is far away from the trend line of recidivism, it proves that the predicted value has errors, and it needs to be further calculated offline through the loss function formula (4), and then the obtained results are analyzed according to the 0.5 threshold. When the predicted value has a small deviation near the trend line, it indicates that the collected data has small float ability, so the error can be ignored. When the predicted value and trend line increase more than 0.5 synchronously, the predicted value from April to June in Figure 7 indicates that the recidivism rate of the person is getting higher and higher, and there is a crime trend. The public security department should focus on offline monitoring for the months exceeding 0.5. If they decrease, the result will be the opposite.

When the predicted value and trend line decrease more than 0.5 synchronously and the deviation between the predicted value and trend line is large, it indicates that the collected data has strong floatability, such as the predicted value from March to June in Figure 8. At this time, it is necessary to calculate the loss function offline to obtain the real predicted value, which will be further analyzed and discussed.

6.3. Result Optimization. The loss function is used to optimize the results shown in Figure 8. The loss function belongs to the error generated by the prediction model in the process of prediction, which indicates that the result predicted by the model is not consistent with the real situation because there are many uncertainties, such as the deviation of the data collection source, the nonstandard processing, and the uncertainty of the synthesis standard during the regression, which may produce errors. The gradient descent algorithm is a widely used optimization algorithm in machine learning. The current popular machine learning library or deep learning library includes different variants of the gradient descent algorithm, and the Sklearn library used in this paper is no exception.

The gradient descent algorithm includes three methods: batch gradient descent, SGD (stochastic gradient descent), and minibatch gradient descent. This research mainly uses SGD to optimize the loss function. SGD is a simple but effective method, which is mostly used for convex loss function optimization in the case of the support vector machine and linear regression. At this time, a point is randomly selected from the predicted points in Figure 8 for gradient descent instead of parameter iteration after traversing all the samples. After training, the gradient is updated once, and then a point is randomly selected for gradient descent and then updated once. After repeated iterations, a model with an acceptable loss value can be obtained. The randomness here means that the samples will be randomly scrambled in each iteration, which can effectively reduce the problem of parameter update cancellation caused by samples.

In the simulation experiment, the data processing and feature engineering modules have been completed, and the SGD regressor method in the Sklearn library is called...
to achieve error optimization to obtain the optimization results, as shown in Figure 9. Compared with the previous linear regression prediction without optimization, the predicted value from April to June obviously fits the predicted line, which indicates that the accuracy of the prediction has been improved. The predicted value gradually decreases from 0.5 in April to 0.35 in June, which is lower than the threshold value of 0.5, indicating that the crime trend of the released prisoners during this period is small, and the public security department can reduce the supervision.

7. Discussions

This research needs to establish a unified platform of multi-source heterogeneous data information. Run the input and output component methods of elastic technology stack to collect and store data, and keep a variety of databases, file systems, and other interfaces in the platform for connecting with the public security, shopping malls, water and electricity bureaus, and other relevant departments to obtain data. This experiment uses the form of simulation to simulate the real released prisoners’ population information, water,
electricity, and gas information, medical treatment information, hotel information, travel information, bank account statement information, payment information, work information, telecommunication data, network data, and other data resources and input them into the prediction model. The Pandas module is used for data processing, including invalid data filtering (filter symbol), missing value filling (missing data supplement), data conversion (data format conversion), and data deduplication (deletion of same data). After processing, the data is segmented (divided into a training set and a test set), and then the dimensionless method (standardized or normalized) and feature engineering are carried out for the segmented data. Feature engineering needs feature extraction (text into data), feature dimensionality reduction (low variance feature filtering), feature statistics (statistics of related data times), etc., which need to be transported to iterator and converter methods. Finally, the linear regression algorithm is used to calculate, and the predicted data is used for analysis and evaluation. The predicted results are presented in the form of a scatter plot, and the predicted value of ordinate 0.5 is the threshold. When the trend line of recidivism rises and the predicted value of a certain month coincides with the trend line of recidivism, which exceeds 0.5, it is judged that there is a trend of recidivism. On the contrary, if the predicted value does not exceed 0.5, it will be judged that there is no trend of recidivism or the possibility of recidivism is low. At this time, the public security department can relax the supervision of the released prisoners. If the original data collected fluctuates greatly and is obviously inconsistent with the predicted value in recent months, the error cannot be ignored, so further calculation should be carried out through the loss function offline to get the predicted value. The linear regression algorithm is not good for overfitting or underfitting, so the algorithm should be improved to make the predicted data more accurate.

The linear regression algorithm is to get the unknown results from the known data, and the applicable data type is numerical. Advantages are as follows: the results have good interpretability, and the calculation of entropy is not complex. Disadvantages are as follows: the nonlinear data fitting is not good. The linear regression algorithm is closely related to data floating. The smaller the data floating is, the smaller the loss function value is, and the more accurate the prediction value will be. On the contrary, if the data floating is large, the resulting error will be large, and the predicted value will not be accurate at this time. Only by calculating the loss function can the more accurate predicted value be obtained. This simulation experiment only adopts one algorithm for predictive analysis, and no other algorithm is used for comparison. Therefore, it cannot effectively highlight which data analysis algorithm can be better used to predict the recidivism of released prisoners, which is also the deficiency of the experiment.

8. Conclusions

This research constructs a prediction model of criminal behavior of specific populations, which includes four parts: data preparation, data segmentation, feature engineering, and evaluation and prediction to meet the practical demands of public security organs, and it uses the data exchange technology of the heterogeneous database to complete data collection and database establishment and uses the feature engineering technology to analyze the big data of specific personnel to obtain multidimensional behavior trajectory data and carry out sorting and statistics. On this basis, the linear regression algorithm is applied to predict and evaluate, and the results are visualized to assist the research and judgment of the possibility of recidivism of specific populations.

![Figure 9: Optimization of prediction results of recidivism of B from January to June 2020.](image-url)
The model programming and simulation experiments are carried out based on Python, and the collection and processing analysis of 20 multidimensional behavior data of two people in the first half of 2020 are completed. The trend chart of crime prediction is obtained, and the loss function is used to optimize the results so as to improve the intuition and accuracy of prediction. The experimental results can help the research and judgment personnel to effectively supervise the specific population to a certain extent to reduce the possibility of recidivism, reduce the workload of the staff, and improve the work efficiency and effect. In the next step, the real test will be carried out to help the public security work in China and contribute to the maintenance of national and social stability.

Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding authors on reasonable request.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Acknowledgments

This work is partially supported by the Research Project of Hubei University of Police and the Key Programs of Education Science and Planning of Hubei Province in 2020 (No. 2020GA049).

References


