Research Article

Prediction of Telecommunication Network Fraud Crime Based on Regression-LSTM Model

Youyang Gao,¹ Dechun Yin,¹ Xiaoliang Zhao,¹ Yu Wang,¹ and Yan Huang²

¹College of Information Network Security, People’s Public Security University of China, Beijing 10038, China
²China Mobile Information Technology Company Limited, China Mobile, Shenzhen 518048, China

Correspondence should be addressed to Yan Huang; huangyan@chinamobile.com

Received 25 May 2022; Revised 12 July 2022; Accepted 20 July 2022; Published 8 August 2022

Copyright © 2022 Youyang Gao 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.

Telecommunication network fraud crimes frequently occur in China. Predicting the number and trend of telecommunication network fraud will be of great significance to combating crimes and protecting the legal property of citizens. This paper proposes a combined model of predicting telecommunication network fraud crimes based on the Regression-LSTM model. First, we find that there is a strong correlation between privacy data illegally sold on the dark web and telecommunication network fraud data. Hence, this paper constructs a Linear Regression model using the privacy data illegally sold on the dark web to predict the number of telecommunication network fraud crimes. Second, an LSTM prediction model is constructed using the data of telecommunication network fraud cases on China Judgments Online based on the time-series feature of telecommunication network fraud crimes. Third, this paper uses the error reciprocal method to combine the two models for prediction. In addition, this paper selects the monthly data set of telecommunication network fraud occurring in 2021 for experimental evaluation. The experimental results show that the accuracy of the Regression-LSTM model constructed in this paper is 86.80%, and the RMSE is 0.149. Compared with the ARIMA, Linear Regression, LSTM, Additive-ARIMA-LSTM, and Multiplicative-ARIMA-LSTM models, the Regression-LSTM model proposed has the highest prediction accuracy.

1. Introduction

Telecommunication network fraud refers to the crime of setting up traps, implementing remote fraud on the victim, and inducing the victim to transfer money through text messages, telephone, and network. Telecom fraud crimes recently have a progressively increasing tendency in China, causing much property damage to people. Thus, it is necessary to take measures, for example, predicting the number of telecom fraud crimes, to curb the trend of the high incidence of telecom fraud. It is of great significance to help the public security departments implement crisis early warning and prevention. However, no one has yet studied the prediction of the number of telecom fraud crimes.

This paper is aimed at building a model of predicting telecommunication fraud by using big data, e.g., the data of personal privacy traded on the dark web and the data of telecommunication network fraud cases on China Judgments Online. The data of personal privacy traded on the dark web is essential to building the predicting model. For criminals, obtaining accurate personal information about the target group is a prerequisite before they commit telecommunication network fraud. Nowadays, the illegal trading of such personal information is mainly made in trading forums on the dark web [1, 2]. Therefore, we should consider the data of personal privacy traded on the dark web as an essential source for building the predicting model of telecom fraud crimes. Furthermore, the occurrence of telecom fraud crimes has its own timing characteristics, so we also use the data of telecommunication network fraud cases on China Judgments Online to analyze the quantitative characteristics of crimes over time.

Hence, this paper builds a combined model (i.e., Regression-LSTM) by using the Linear Regression and the LSTM model to predict the number of telecom fraud crimes. The Linear Regression submodel is used for finding the causality feature and correlation between the trading data of the dark web and the crimes and then predicting the number of
telecom fraud crimes. In addition, the LSTM submodel is used for mining the time-series feature of the crimes and predicting the quantitative characteristics with the monthly data set of telecommunication fraud cases on China Judgments Online.

Different from general prediction models only based on the time-series feature, the Regression-LSTM model considers both the causality feature and the time-series feature of telecom fraud crimes. This paper uses the error reciprocal method to make a weighted combination of the Linear Regression submodel and the LSTM submodel. The Regression-LSTM model is compared with the ARIMA model, Linear Regression model, LSTM model, Additive-ARIMA-LSTM model, and Multiplicative-ARIMA-LSTM model, and it obtains the best prediction effect.

2. Related Works

At present, the academic research on the crime of telecommunication network fraud mainly focuses on the causes, prevention and control measures [3–5], and governance schemes [6, 7]. Few scholars study the prediction of the number and trend of telecommunication network fraud. Currently, crime prediction is more focused on crimes except telecom fraud. For example, Yu et al. [8] predict the trend of juvenile delinquency based on the Grey prediction method. Yu et al. [9] use a fuzzy BP neural network to predict the number of crimes in Dalian over the years 2000-2009. Feng et al. [10] use the Prophet model and the LSTM model to predict the number of crimes in three major cities in the United States. Chen et al. [11] use the Gray-Markov model to analyze the weekly crime rate of property intrusion in Langfang City from January to June 2013. Tu and Chen’s study [12] is based on the ARIMA-LSSVM mixed model for the combined prediction of small sample crime time series. Jha et al. [13] propose an improved ARIMA model and an improved ANN model to predict the number of crimes in India in the next five years. Zhai et al. [14] use the Prophet model to predict the number of crimes in a long-term time series. Shen et al. [15] use LSTM to construct an optimal binary alarm data model and frequency statistical data model to predict the probability and number of crimes in each region of a city in China. Hossain et al. [16] predict the number of crimes in San Francisco from 2003 to 2015 based on the Adaboost improved random forest algorithm. Yan and Hou [17] predict the changing trend of the number of daily theft crimes based on the LSTM model. Liu [18] predicts the number of crimes in large cities based on an improved LSTMNet model.

The above prediction methods for various crimes are mainly based on the time-series data to predict the risk of crimes in a future period. The influence of factors besides time on criminal activity is not considered.

Compared with the above studies only based on time-series features of crimes, this paper combines the causality feature and the time-series feature of telecom fraud crimes. This paper finds a strong correlation between personal privacy data traded on the dark web and telecommunication network fraud crimes. On this basis, a Linear Regression prediction model is built for telecommunication network fraud crimes. In addition, we use the real telecom fraud cases from January 2014 to December 2021 on China Judgments Online to build an LSTM network model that can effectively learn the time-series feature of telecom fraud crimes. We further build a Regression-LSTM model based on these two models with the error reciprocal method. Regression-LSTM can fully combine the respective advantages of the Linear Regression prediction model and the LSTM prediction model to achieve better prediction results.

3. Introduction to Experimental Data

The data used in this paper mainly includes personal privacy data traded on the dark web and judgment document data of telecom fraud published on China Judgments Online. Personal privacy data traded on the dark web is a critical data source. It is an important channel for public security departments to investigate and combat crimes. The judgment document data of telecom fraud published on China Judgments Online is real telecommunication fraud case data solved by the police departments.

3.1. Personal Privacy Data on the Dark Web

First, we enter the dark web with specific technical means and choose the most popular Chinese trading forum containing multiple trading sections on the dark web as our data source. Next, we find out that the personal privacy data sold in the section “Data Resources” is closely related to the crime of telecommunication network fraud. So we use crawlers to obtain the whole data for this section. From June 2018 to June 2021, there are 4,340 posts related to personal privacy in this section. The statistical figure of the number of new posts on the dark web per month is shown in Figure 1. As can be seen from the figure, the privacy data leaked on the dark web every month has an obvious cyclical fluctuation in quantity.

3.2. Telecommunication Network Fraud Data on China Judgments Online

China Judgments Online (https://wenshu.court.gov.cn/) publishes information on various cases that have been solved so that we can inquire about cases related to telecommunication network fraud. According to China Judgments Online, there were 9,190 judgment documents related to telecommunication network fraud from 2010 to 2021. Among the judgment documents, there are 332 in connection with fake prosecutor fraud, 288 about feelings cheating fraud, 227 about fake customer service fraud, 181 about abnormal online shopping fraud, 54 about fake boss fraud, 67 about fake teacher fraud, 2,420 about loan fraud, 449 about fake credit card fraud, 43 about fake ticket fraud, 213 about fake acquaintance fraud, 62 about game fraud, 19 about naked chat blackmail, and 4,835 about various other forms of judgment documents for telecommunication network fraud.

The number distribution of telecom network fraud cases is shown in Figure 2. It can be seen from the figure that from 2014 to 2017, the number of telecommunication network fraud crimes in China was relatively small. After 2018, telecommunication fraud crimes surged and showed periodic
fluctuations. This paper selects the data of telecommunication fraud crimes on China Judgments Online from January 2014 to December 2021 for time series prediction.

4. Prediction of Telecom Fraud Crime Based on Regression-LSTM

Due to the strong correlation between the personal privacy data sold on the dark web and telecommunication network fraud crimes, we establish a Linear Regression model to predict telecommunication network fraud crimes. Since the criminal case data is time-series data, it is possible to conduct time-series modeling and predict the number and trend of future crimes. Therefore, we also establish an LSTM model to predict telecommunication network fraud crimes. For the above two models, although the Linear Regression model can predict the number of telecom fraud crimes based on the causality feature, the model does not consider the time-series feature of the crimes themselves. Meanwhile, although the LSTM model can predict the number of telecom fraud crimes based on the time-series feature, it ignores the causality feature of the crimes. Therefore, this paper further uses the error reciprocal method to perform the weighted summation of the two models, constructing a telecom fraud prediction model based on the combined Regression-LSTM model. The architecture of this combined model is shown in Figure 3.

Based on the dark web privacy data and telecom fraud data obtained by crawlers, this paper firstly constructs a correlation matrix between the two. It then uses Spearman correlation analysis to verify that the two data are strongly

---

**Figure 1:** Number of new posts on the dark web per month.

**Figure 2:** Number distribution of telecommunication fraud cases.
correlated. Because of the strong correlation between the two kinds of data, this paper trains a Linear Regression model to predict telecommunication fraud after normalizing them. Then, based on the time-series features of telecom fraud data, we standardize the telecom fraud data and calculate the loss function. An LSTM telecom fraud prediction model is constructed with the Adam optimization algorithm and MMSE (Minimize Mean Square Error) principle. Finally, we combine these two models with the error reciprocal method to construct a Regression-LSTM model to predict the monthly number of telecom fraud cases in 2021.

4.1. Prediction of Telecom Fraud Based on Linear Regression

4.1.1. Correlation Matrix. According to the announcement of the Beijing Anti-Telecom Network Fraud Center, there are currently twelve main types of telecom fraud crimes: fake prosecutor fraud, feelings cheating fraud, fake customer service fraud, abnormal online shopping fraud, fake boss fraud, fake teacher fraud, loan fraud, fake credit fraud, fake ticket fraud, fake acquaintance fraud, game fraud, and naked chat blackmail. Each kind of transaction data on the dark web may be used for telecom fraud. For example, loan data can be used for loan fraud; student data can be used to fake teacher fraud; shopping data can be used for abnormal online shopping fraud. We study the criminal process of these twelve types of telecom fraud and manually analyze the posts of privacy data transactions on the dark web. Each transaction post on the dark web contains content that can be used for some fraud case, it is marked as 1. Otherwise, it is marked as 0. Based on this, a correlation matrix $A$ with $m$ rows and $n$ columns is constructed:

$$A = \begin{bmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} \\
  a_{21} & a_{22} & \cdots & a_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix},$$ (1)

where $A$ is a 0/1 matrix, $a_{ij}$ indicates whether the $i$th post is related to the $j$th fraud type. The associated value is 1. Otherwise, the value is 0. We sample 1000 posts to study the relationship between the posts and the twelve case types. At last, we construct a matrix $A_{1000,12}$ of 1000 rows and 12 columns, as shown in Figure 4.

After further analysis of the data sold by each post in the forum, we find that the number of data records sold by different posts is often different. For example, in a post titled “83W personal information_rich_manager, etc.” the amount of privacy records it sells is 830,000, although there is only one post. Therefore, the number of privacy records involved in the post should also be considered when constructing the correlation matrix. On this basis, we continue to improve the correlation matrix, defining a correlation matrix $B$ with weights:

$$B = \begin{bmatrix}
  b_{11} & b_{12} & \cdots & b_{1n} \\
  b_{21} & b_{22} & \cdots & b_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  b_{m1} & b_{m2} & \cdots & b_{mn}
\end{bmatrix},$$ (2)

In this matrix, $b_{ij}$ indicates that the $i$th post is related to the $j$th type of case, and the number of privacy records sold by this
As can be seen from Table 1, the amount of privacy data sold on the dark web is very large. In all types of privacy data, the amount of data used for loan fraud is the largest (2.539149668 billion), followed by the data used for naked chat blackmail (1.878473666 billion) and the data of fake prosecutor fraud (1.2756833 billion). Among the judgment documents, loan fraud crimes are the most (2420), followed by the data used for fake credit fraud (449), then fake prosecutor fraud (332), and feelings cheating fraud (288). Obviously, we can find that the amount of privacy data used for loan fraud is the largest, consistent with most loan fraud crimes in the judgment documents. This explains why loan fraud crimes are so common in real life.

In order to further analyze correlations behind these two sets of data, according to Table 1, we have drawn the distribution curve of the number of judgment documents for telecom fraud published on China Judgments Online and the distribution curve of the number of personal privacy data records sold on the dark web, as shown in Figure 6.

Among them, the horizontal axis represents 12 types of telecom fraud, the left vertical axis represents the number of personal privacy data records sold on the dark web, and the right vertical axis represents the number of judgment documents for telecom fraud. As Figure 6 shows, it can be concluded that the distribution of the quantity of data traded on the dark web and the distribution of the number of judgment documents on China Judgments Online are generally consistent. The reason for this may be as follows: Before the occurrence of telecom network fraud, some related data has been sold on the dark web, so criminals are likely to buy this data for precise fraud.

4.1.2. Correlation Analysis between Privacy Data and Telecom Fraud Cases

(1) Comparison between the Number of Judgment Documents for Telecom Fraud and the Amount of Privacy Data Sold on the Dark Web. How is the number of cases published on China Judgments Online distributed and whether there is some similarity or correlation with the distribution of the number of personal privacy data records sold on the dark web? To solve this question, we sum up each column of the weighted correlation matrix separately to calculate the total amount of data that could be used to commit each type of telecommunication network fraud. Furthermore, combined with the number of various types of judgment documents for telecom fraud on China Judgments Online, a numerical comparison study was conducted, as shown in Table 1.

As Figure 6 shows, it can be concluded that the distribution of the quantity of data traded on the dark web and the distribution of the number of judgment documents on China Judgments Online are generally consistent. The reason for this may be as follows: Before the occurrence of telecom network fraud, some related data has been sold on the dark web, so criminals are likely to buy this data for precise fraud.

### Table 1. Number of Judgment Documents and Amount of Privacy Data

<table>
<thead>
<tr>
<th>Post</th>
<th>Fake prosecutor fraud</th>
<th>Feelings cheating fraud</th>
<th>Fake customer service fraud</th>
<th>Abnormal online shopping fraud</th>
<th>Fake boss fraud</th>
<th>Fake teacher fraud</th>
<th>Loan fraud</th>
<th>Fake credit fraud</th>
<th>Fake ticket fraud</th>
<th>Fake acquaintance fraud</th>
<th>Game fraud</th>
<th>Naked chat blackmail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post12</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post13</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post14</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Post15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 4:** Correlation matrix between posts and types of telecom fraud cases.
Since the number of different types of data on the dark web is different, the probability of criminals committing corresponding types of telecom fraud is also different. Additionally, criminals will keep buying such privacy data if they find that data-backed frauds are easier to pull off. At the same time, more and more such privacy data will be transported to the dark web through the underground black industry chain. In the end, a mutual causality cycle is formed, making illegal privacy data transactions and telecom fraud on the dark web expand viciously.

It is worth noting that the distribution on the right side of the two curves has opposite trends. That is to say, the number of cases of game fraud and naked chat blackmail published on China Judgments Online is relatively small, while the number of corresponding types of data transactions on the dark web is relatively large.

For game fraud, there is a lot of such privacy data sold on the dark web. Accordingly, there are many potential defrauded people. However, the number of judgment documents about such cases is relatively small. The reasons may be as follows: First, from the perspective of victims, the majority of people participating in online games are minors. This group lacks legal awareness. Once they encounter game fraud, most of them will not choose to call the police. Secondly, in terms of family supervision, many teenagers will be supervised by their parents in the financial investment of online games, thus reducing the risk of being defrauded. Thirdly, game fraud crimes are inherently difficult to detect. Many games in the market are not strictly regulated, and there are few effective ways to find criminals, resulting in a low crime-solved rate. Fourthly, the force of police departments is generally insufficient at present, and the work intensity of police is very high.

For naked chat blackmail, there is also a lot of privacy data, such as "single dating" and "hotel accommodation," sold on the dark web, which may be used for naked chat blackmail. However, the number of judgment documents in such cases is relatively small. The main reason may be that...
many victims constrained by reputation are always unwilling to call the police.

(2) Correlation Analysis Based on Spearman. Based on the preliminary statistical comparison and manual qualitative analysis, the above discussion proves that the correlation between posts and telecom fraud crimes exists. Next, by the Spearman correlation analysis, we calculate the correlation coefficient to obtain an accurate measure of the degree of the correlation.

Correlation analysis refers to the analysis of the degree of correlation and closeness between two or more variables. The correlation coefficient is a measure of correlation analysis. The value of the correlation coefficient is between [-1, +1], which represents the degree of association between two random variables [19]. When the correlation coefficient is less than 0, it is called negative correlation; when it is greater than 0, it is called positive correlation; when it is equal to 0, it is called zero correlation. The larger the absolute value of the correlation coefficient, the stronger the correlation between the two variables. The closer the correlation coefficient value is to -1 or +1, the closer the relationship is; when it is close to 0, the relationship is distant. Generally speaking, when the absolute value of the correlation coefficient is less than 0.2, it indicates a very weak correlation or no correlation. When it is between 0.2 and 0.4, it indicates a weak correlation. When it is between 0.4 and 0.6, it indicates a moderate degree of correlation, and when it is between 0.6 and 0.8, it indicates a strong correlation, and when it is greater than 0.8, it indicates a very strong correlation.

The commonly used correlation analysis methods mainly include Spearman and Pearson. Spearman is a non-parameter analysis method that is independent of the data distribution. But when we use Pearson for correlation analysis, the sample data must obey the normal distribution [20]. To choose a suitable correlation analysis method, we first test the normality of the sample data.

In this paper, with SPSS 25.0, we use Kolmogorov-Smirnov single-sample normality test method to test the normality of the sample data. The results are shown in Table 2.

As can be seen from Table 2, the significance level of the sample data is less than 0.05, meaning that the sample data does not obey the normal distribution. Therefore, Spearman
Table 3: Spearman correlation analysis.

<table>
<thead>
<tr>
<th>Correlation coefficient of Spearman</th>
<th>Posts on the dark web</th>
<th>Telecom fraud cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posts on the dark web</td>
<td>1.000</td>
<td>0.782**</td>
</tr>
<tr>
<td>Telecom fraud cases</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>

**The correlation is significant at the 0.01 level.

is chosen to verify the correlation between data on the posts on the dark web and telecom fraud cases.

The principle of Spearman correlation analysis is as follows [21]:

\[ \sigma_{xy} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \]  

where \( \mu_x \) and \( \mu_y \) are the mean of \( X \) and \( Y \), respectively, and \( d_i = x_i - y_i \) is the result of calculation after grading the variables separately.

Before calculating the correlation between the number of various telecom fraud crimes and the number of corresponding posts on the dark web, according to the above discussion, game fraud and naked chat blackmail are excluded. The results of the Spearman correlation analysis are shown in Table 3.

As can be seen from Table 3, the correlation coefficient between the amount of privacy data sold on the dark web and the number of different types of fraud cases is 0.782, which is at a strong correlation level, indicating that there is a strong correlation between the posts on the dark web and telecom fraud cases.

4.1.3. Prediction of the Number and the Trend of Telecommunication Fraud. As concluded above, because of the strong correlation between the privacy data on the dark web and telecommunication fraud data, we can predict the number of telecommunication fraud cases in the future based on the number of privacy data sold on the dark web.

(1) The Interval between Posting Time and Case Closing Time. First of all, it usually takes a while for criminals to commit telecom fraud, and it will also take a while for the public security department to solve the telecom fraud crimes after the calling. Therefore, compared with the time when the privacy data transaction on the dark web occurs, there must be a certain lag when the judgment document is published online after the case is solved. If we want to predict the number of telecom fraud crimes, we need to calculate the general interval.

From August 2018 to December 2021 (41 months), the monthly number of posts on the dark web and the monthly number of judgment documents for telecommunication fraud on China Judgments Online are counted. Part of the data is shown in Table 4. We define two vectors, \( X \) and \( Y \), representing the set of posts on the dark web and the set of judgment documents, respectively.

\[ X = \{x_1, x_2, \ldots, x_0, \ldots, x_{12}\} \], where \( x_i \) indicates the number of posts posted on the dark web in the \( i \)th month of 2019. The dimension of \( X \) is 12, which indicates that \( X \) is a dataset of posts for 12 consecutive months on the dark web.

\[ Y = \{y_{1+\Delta t}, y_{2+\Delta t}, \ldots, y_{12+\Delta t}\} \], where \( y_{i+\Delta t} \) indicates the number of telecom fraud documents on China Judgments Online in the \( (i + \Delta t) \)th month. The dimension of \( Y \) is 12, which indicates that \( Y \) is a data set of judgment documents for telecommunication fraud on China Judgments Online for 12 consecutive months. The value range of \( \Delta t \) is the integer interval \([0, 17] \). When \( \Delta t \) takes values from small to large in \([0, 17] \), a set of vectors \( Y_0, Y_1, Y_2, \ldots, Y_9, Y_{10}, \ldots, Y_{17} \) will be obtained. \( Y_0 \) is the data set of judgment documents in the same year and month as \( X \), \( Y_{17} \) is the data set of judgment documents shifted by \( j \) months from the month where \( X \) is located.

We sequentially calculate the correlation between \( X \) and each vector of \( Y_0, Y_1, Y_2, \ldots, Y_9, Y_{10}, \ldots, Y_{17} \), obtaining 18 sets of Spearman correlation coefficient values, as shown in Table 5.

The largest Spearman correlation coefficient is 0.783, and the corresponding \( \Delta t \) is 13, which means that when \( Y \) is 13 months later than \( X \), and the Spearman correlation coefficient between \( X \) and \( Y \) is the largest. As a result, we infer that the interval from the time when posts appear on the dark web to the time when the telecom fraud case is solved is about 13 months.

(2) Linear Regression Prediction Model. This paper uses a Linear Regression model to predict the number of telecom fraud cases. First, we construct a training data set with the number of posts on the dark web from August 2018 to November 2019 and the number of telecom fraud cases 13 months later (from September 2019). Then, we use the least-squares method to perform linear fitting on the training data set and complete the model training after obtaining the parameter values in the Linear Regression model. The specific process is as follows:

We define a vector \( X = \{x_1, x_2, \ldots, x_0, \ldots, x_6\} \), where \( X \) represents a data set consisting of the number of posts on the dark web for 16 consecutive months, and \( x_i \) indicates the number of posts in the \( i \)th consecutive month. The value of \( X \) is as follows:

\[ X = \{61,209,236,207,112,120,53,114,161,201,232,233,300,198,127,0\} \].

(6)

We define a vector \( Y = \{y_{1+\Delta t}, y_{2+\Delta t}, \ldots, y_{16+\Delta t}\} \), where \( \Delta t = 13 \); thus, \( Y = \{y_{14}, y_{15}, \ldots, y_{29}\} \), where \( Y \) represents the data set of judgment documents for telecommunication fraud for 16 consecutive months from September 2019. The value of \( Y \) is as follows:


(7)
Table 4: Number of monthly posts on the dark web and monthly fraud cases from 2018 to 2021.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of posts in 2018</td>
<td>61</td>
<td>209</td>
<td>236</td>
<td>207</td>
<td>112</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of posts in 2019</td>
<td>120</td>
<td>53</td>
<td>114</td>
<td>161</td>
<td>201</td>
<td>232</td>
<td>233</td>
<td>300</td>
<td>198</td>
<td>127</td>
<td>0</td>
<td>86</td>
</tr>
<tr>
<td>Number of posts in 2020</td>
<td>70</td>
<td>106</td>
<td>143</td>
<td>105</td>
<td>109</td>
<td>78</td>
<td>58</td>
<td>72</td>
<td>70</td>
<td>69</td>
<td>40</td>
<td>86</td>
</tr>
<tr>
<td>Number of telecom fraud cases in 2019</td>
<td>162</td>
<td>99</td>
<td>188</td>
<td>210</td>
<td>196</td>
<td>209</td>
<td>207</td>
<td>202</td>
<td>179</td>
<td>256</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>Number of telecom fraud cases in 2020</td>
<td>113</td>
<td>160</td>
<td>115</td>
<td>162</td>
<td>250</td>
<td>201</td>
<td>232</td>
<td>280</td>
<td>195</td>
<td>238</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>Number of telecom fraud cases in 2021</td>
<td>200</td>
<td>214</td>
<td>220</td>
<td>255</td>
<td>172</td>
<td>196</td>
<td>252</td>
<td>172</td>
<td>196</td>
<td>252</td>
<td>172</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Spearman correlation coefficient when the month interval between X and Y is Δt.

<table>
<thead>
<tr>
<th>Δt (month interval between X and Y)</th>
<th>Spearman correlation coefficient between X and Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.371</td>
</tr>
<tr>
<td>1</td>
<td>0.357</td>
</tr>
<tr>
<td>2</td>
<td>0.319</td>
</tr>
<tr>
<td>3</td>
<td>0.552</td>
</tr>
<tr>
<td>4</td>
<td>0.252</td>
</tr>
<tr>
<td>5</td>
<td>-0.210</td>
</tr>
<tr>
<td>6</td>
<td>-0.538</td>
</tr>
<tr>
<td>7</td>
<td>-0.706</td>
</tr>
<tr>
<td>8</td>
<td>-0.713</td>
</tr>
<tr>
<td>9</td>
<td>-0.713</td>
</tr>
<tr>
<td>10</td>
<td>-0.147</td>
</tr>
<tr>
<td>11</td>
<td>0.266</td>
</tr>
<tr>
<td>12</td>
<td>0.524</td>
</tr>
<tr>
<td>13</td>
<td>0.783</td>
</tr>
<tr>
<td>14</td>
<td>0.769</td>
</tr>
<tr>
<td>15</td>
<td>0.573</td>
</tr>
<tr>
<td>16</td>
<td>0.224</td>
</tr>
<tr>
<td>17</td>
<td>0.077</td>
</tr>
</tbody>
</table>

In the above formula, \(a = \bar{y} - \bar{x}b = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}\).

The training set \([X_{\text{norm}}, Y]\) is constructed according to \(X_{\text{norm}}\) and \(Y\), and the training result is shown in Figure 8.

The RMSE (Root Mean Square Error) of training data is 32, and the \(R^2\) is 0.63, indicating that the regression model fits well. \(P = 0.01 < 0.05\), indicating that the results of the model are reliable.

The parameters obtained by training \(a = 160.4\), and \(b = 110.3\), so the Linear Regression model is as follows:

\[Y = 160.4 + 110.3X_{\text{norm}}.\] (10)

We use the trained Linear Regression model to make predictions. This paper predicts the number of telecommunication fraud cases to be closed each month in 2021 based on the number of monthly posts on the dark web from December 2019 to November 2020.

We define a data set \(X = \{x_1, x_2, \ldots, x_{12}\}\) that represents the number of posts on the dark web from December 2019 to November 2020, where \(x_i\) represents the number of posts on the dark web for the \(i\)th consecutive month since December 2019. The value of \(X\) is as follows:

\[X = \{86, 70, 106, 143, 105, 109, 78, 58, 72, 70, 69, 40\}.\] (11)

The number of telecom fraud cases in all months in 2021 predicted by this paper based on the Linear Regression model is represented by the vector \(Y_{\text{regression}}\). We calculate \(Y_{\text{regression}}\) based on \(X\) as follows:

\[Y_{\text{regression}} = \{206, 188, 229, 270, 228, 232, 197, 175, 191, 188, 187, 155\}.\] (12)

4.2. Prediction of Telecom Network Fraud Crime Based on LSTM Model. Time series prediction of crimes is one of the main directions of criminology. The time-series model can be used to analyze the statistical laws of time-series data and further predict the number of telecom fraud crimes. This paper uses the Long Short-Term Memory (LSTM) model to predict the time series of telecom fraud crimes.

4.2.1. Introduction to LSTM Model. Long Short-Term Memory (LSTM) is a Recurrent Neural Network (RNN) variant. Theoretically, RNN can handle any long-distance dependency problem [22, 23]. However, due to its gradient disappearance, gradient explosion, and other problems, RNN has
only short-term memory and cannot achieve long-term preservation of information. Therefore, LSTM solves the long-term dependence of information by adding internal gating mechanisms and memory cells to maintain the long-term preservation of information. The LSTM cell structure is shown in Figure 9.

The LSTM unit includes an input gate $i_t$, a forget gate $f_t$, and an output gate $o_t$. The input and output vectors of the LSTM hidden layer are $x_t$ and $h_t$, and the memory unit is $C_t$. The input gate is used to control the amount of current input data $x_t$ flowing into $C_t$, that is, the amount of input information saved to $C_t$. The forget gate controls the retention and forgetting of information and avoids the gradient disappearance problem caused by the back-propagation of the gradient over time in a specific way, that is, the influence of the previous moment information $C_{t-1}$ on the current moment $C_t$. The output gate can control the influence of memory unit $C_t$ on the current output value $h_t$. It controls the output information of memory unit $C_t$ at time $t$. The output $h_t$ is not only affected by $C_t$ and $x_t$, but also by $h_{t-1}$. Memory unit $C_t$ is not only affected by $h_{t-1}$ and $x_t$, but also by $C_{t-1}$.

At time $t$, the mathematical expression of each gate state is as follows:

$$C_t = \tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c),$$
$$i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i),$$
$$f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f),$$
$$o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o),$$
$$C_t = f_t \otimes C_{t-1} + i_t \otimes C_t,$$
$$h_t = o_t \otimes \tanh(C_t),$$

where $\sigma$ is the sigmoid activation function, and $\otimes$ is the Hadamard product.
where \( i \) is the intermediate calculation result of the input information and past information passing through the input gate and \( f \) is the intermediate calculation result of the input information and past information passing through the forget gate. \( o \) is the intermediate calculation result of the input information and past information passing through the output gate. \( C \) is the intermediate result of the input information and past information passing through the tanh activation function, \( \sigma \) is the sigmoid activation function, \( \otimes \) represents the dot product between elements, multiplied point by point, \( W \) and \( U \) are the weight matrices of the neural network, and \( b \) is the paranoid vector of the neural network.

4.2.2. Test of LSTM Model. This model uses the data of 96 months of telecommunication fraud judgment documents from January 2014 to December 2021. The data on telecommunication fraud occurring during the first 84 months is used as the training set, and the data occurring during the last 12 months is used as the test set.

The stride of the time series of LSTM is 24, and the label length is 12. This model uses the min/max normalization method. The normalized data range is \([-1, 1]\). After the data is fed into the LSTM model, the loss function is calculated. In the model training process, the Adam algorithm is used to optimize, and the solution is solved based on the principle of minimum Root Mean Square Error.

The loss function curves of the training set and the validation set of the model training process are shown in Figure 10, and the accuracy curve of the validation set is shown in Figure 11.

![Loss function curve](image)

**Figure 10: Loss function.**

As shown in Figures 10 and 11, with the increase in training times, the accuracy rate also increases, and then, the accuracy rate gradually becomes stable. The loss function of the model begins to converge after 20 cycles, and the model reaches the optimal performance at this time, saving the internal parameters of the model.

This paper repeats the experiment 20 times to predict the number of telecom fraud crimes for all months in 2021 based on LSTM and perform Kolmogorov-Smirnov tests on the obtained results, as shown in Table 6.

As can be seen from Table 6, the significance level of 20 sets of predicted results is less than 0.05, which means that each set conforms to a normal distribution. Therefore, we use the mean of 20 sets of results as the prediction result of the LSTM model. It is represented by the vector \( Y_{\text{LSTM}} \):

\[
Y_{\text{LSTM}} = [182, 182, 165, 156, 151, 159, 169, 189, 218, 231, 240, 239].
\]

4.3. Hybrid Prediction Models. The Linear Regression model predicts the number of telecommunication fraud crimes based on the correlation between privacy data and telecommunication fraud. At the same time, the LSTM model studies the time-series feature of telecom fraud crimes to make the prediction. This paper proposes a Regression-LSTM combination model, combining the advantages of the two models to predict the number of telecom fraud crimes.

First, the errors of the Linear Regression model and the LSTM model are calculated separately. Then, the prediction results of the two models are weighted and combined with the error reciprocal method to the final prediction result \( f_{RL} \). The principle of the error reciprocal method is as follows:

\[
f_{RL} = \omega_1 f_R + \omega_2 f_L,
\]

\[
\omega_1 = \frac{e_2}{e_1 + e_2},
\]

\[
\omega_2 = \frac{e_1}{e_1 + e_2},
\]

where \( \omega_i \) represents the weight coefficient of the model,
\( f_R \) is the predicted value of the Linear Regression model, \( f_L \) is the predicted value of the LSTM model, \( e_1 \) is the error of the Linear Regression Model, and \( e_2 \) is the error of the LSTM model. According to (15)–(17), it can be seen that through the error reciprocal method, the two models can be combined to reduce the overall error. Thus, the Regression-LSTM model can improve the prediction accuracy.

In addition, the convergence of the Regression-LSTM model is based on the Linear Regression model and the LSTM model, respectively. Since the Linear Regression model and the LSTM model are convergent, the Regression-LSTM model, a linear combination of the two models, is convergent.

In order to compare the Regression-LSTM model with existing hybrid models, this paper constructs the Additive-ARIMA-LSTM hybrid model and Multiplicative-ARIMA-LSTM hybrid model for time series prediction of telecom network fraud crimes [24].

In the Additive-ARIMA-LSTM method as in Figure 12, the time series \( y = [y_1, y_2, \ldots, y_n]^T \) is considered as an addition of a linear (L) and a nonlinear (N) component as Equation (18). First, a linear model is applied on the time series to obtain the forecasts on linear component (L). Then, the residual series (e) is computed by subtracting the forecasts on linear component (L) from the original time series y as Equation (19). The residual series is used by a nonlinear model to obtain the forecasts on nonlinear component N. Then, the final forecasts are obtained by adding the forecasts...
on linear component with the forecasts on nonlinear component as in Equation (20). In this paper, we have used ARIMA as the linear model and LSTM as the nonlinear model. Hence, Additive-ARIMA-LSTM is obtained and used for forecasting.

\[
y = L + N, \quad \text{(18)}
\]

\[
e = y - \hat{N}, \quad \text{(19)}
\]

\[
\hat{y} = \hat{L} + \hat{N}. \quad \text{(20)}
\]

5. Evaluation of the Models

This paper uses the measurements of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), accuracy (ACC), and CPU time to compare the performance of different models. MAE is the average value of absolute error, which is a primary measurement of error. MAPE can examine the ratio of error to actual value based on MAE. RMSE is the ratio of the square of the deviation between the model predicted value and the real value to the number of predictions \(n\). It is susceptible to significant or minor errors in the prediction data, so it can better reflect the accuracy of the prediction results. ACC can reflect the accuracy of the prediction results and is an essential measurement for evaluating the model. CPU time describes the time consumption of the algorithm.

\[
\text{MAE} = \frac{\sum_{i=1}^{n} |X_{\text{obj},i} - X_{\text{model},i}|}{n},
\]

\[
\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|X_{\text{obj},i} - X_{\text{model},i}|}{X_{\text{obj},i}}, \quad \text{(24)}
\]

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (X_{\text{obj},i} - X_{\text{model},i})^2}{n}},
\]

where \(X_{\text{obj},i}\) is the real value; \(X_{\text{model},i}\) is the predicted value of the model; \(n\) is the number of predictions.

\[
\text{Accuracy} = \left(1 - \sqrt{\frac{1}{n} \sum_{i=1}^{n} E_i^2}\right) \times 100\%, \quad \text{(25)}
\]

where \(n\) is the total number of predicted \(X\) and \(E_i\) is the relative error of a certain point and is defined as follows:

\[
E_i = \frac{|\text{actual} - \text{predict}|}{\text{actual}} \times 100\%. \quad \text{(26)}
\]

This paper compares the Linear Regression Model, the LSTM model, the ARIMA model, the Additive-ARIMA-LSTM model, and the Multiplicative-ARIMA-LSTM model with the Regression-LSTM model, as shown in Figure 14. The MAE, MAPE, RMSE, ACC values, and CPU time of each model are shown in Table 7. The experimental environment is as follows: AMD Ryzen 5 3600 6-Core CPU@3.60 GHZ, 16 G memory, Windows 10, python3.9, torch1.10.0.

It can be seen from Figure 14 that most models have high accuracy in predicting the number of telecommunication network fraud crimes. The trend and the number of fraud crimes predicted by the Regression-LSTM model are almost identical to the actual number of crimes. The Regression-LSTM model has a higher fitting degree and a better prediction effect than other models. Moreover, according to Table 7, the Regression-LSTM model proposed in this paper has the best performance. The MAE and RMSE values of this model are 0.122 and 0.149, smaller
than other single models. The ACC value reaches 86.80%, indicating the highest prediction accuracy. In addition, the CPU time of the Regression-LSTM model is the shortest among the hybrid models.

In Figure 14, the number of telecom fraud crimes approaches 200 almost every month. In March and October, the number is high, which reminds the public security department to pay more attention to the crimes in the two months. Although the curve of real value and the curve of Regression-LSTM do not overlap, their shapes and trends are roughly the same, which shows that it is feasible and practical to predict the amount and trend of telecom fraud crimes based on Regression-LSTM model.

The Regression-LSTM model has many applications in public security. First, we can predict the number distribution of telecom fraud crimes based on the model so that the police department can focus on antifraud publicity during the high-frequency time of crimes and strengthen early warning and preventive measures in a targeted way. Second, as an intelligent analysis tool for crime prediction, this model can be applied to a city to evaluate the rate of fraud crimes and improve its public security level. Last but not least, the model can improve the level of cyberspace security management. If the number of telecom fraud crimes surges in a certain period of time, we can conclude that a large number of personal privacy data were stolen and sold on the dark web. It reminds the internet regulators to track specific virtual identities and crackdown on the underground illegal trading on the dark web.

### 6. Conclusion

In this paper, the Regression-LSTM model is proposed to realize the prediction of telecommunication network fraud crimes. Monthly telecom network fraud data of 2021 on China Judgments Online is selected for model testing. Compared with the other models, the proposed combined model can make up for the insufficiency of a single model and better utilize both the causality features of crimes and the time-series features, reducing the error of a single prediction model. The experimental results show that the combined model Regression-LSTM has the highest prediction accuracy, and the RMSE value is lower than other single prediction models, indicating the best prediction effect.
The Regression-LSTM model combines the advantages of the Regression model and LSTM model. It considers both the causality features of telecom fraud crimes and the time-series features. The Linear Regression submodel is used for finding the causality features and correlation between the trading data of the dark web and the crimes, and the LSTM submodel is used for mining the time-series features of the crimes. This paper confirms that the privacy data sold on the dark web strongly correlates with telecom fraud crimes, providing a new direction for subsequent researchers to research telecom fraud crimes by using the data on the dark web.

The Regression-LSTM model provides a useful analytical tool for telecom fraud prediction. It is of great significance to the protection of public safety. The prediction results of the model can not only remind citizens to strengthen awareness of fraud prevention but also urge enterprises to fulfill their responsibilities of privacy data protection. Under the precise guidance of the prediction model, the police departments can effectively reduce the occurrence of telecom fraud.

In addition, there are still some shortcomings in this paper. In the follow-up research, we will increase the scope and quantity of open-source data to improve the prediction accuracy of telecommunication network fraud crimes.

Data Availability

The dataset used in this study can be provided by the first author and corresponding author upon request. The data are not publicly available due to restrictions, i.e., privacy or ethics.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research was funded by the Special Project of the Ministry of Public Security’s Basic Work of Strengthening Police by Science and Technology “Research on risk group analysis model based on dynamic spatial relationship network” (No. 2020GABJC02) and “Research on the Early Warning Model and Prevention Mechanism of Telecommunication Network Fraud Crime,” Fundamental Research Project of PPSUC (No. 2021JKF420).

References


