Journal of Globe Scientific Reports

journal homepage: www.journal-gsr.online

Paper Type: Original Article

Enhancing Telecom Fraud Prediction Accuracy Using a Combined CNN-LSTM Model with Bahdanau Attention Mechanism

Mingjian Wang[1,*,#], Jingcheng Xie [2,*,#], Zejun Chen[3,*,#], Minghao Liu[4,*,#],

Shaoyang Zhang[5,*,#], Jidong Li[6,*,#]

1.School of Information Science & Engineering, Yunnan University, Kunming, China

2.School of Software Engineering, Tongji University, Shanghai, China

3.Faculty of Social Sciences, University of Macau, Macao, China

4.College of Water&Architectural Engineering, Shihezi University, Shihezi, China

5.College of Physical Science and Technology, Xiamen University, Liaoning Province, Xiamen,China

6.Economics College,Changchun University Of Finance And Economics,changchun, China

#These authors contributed equally to this work and should be considered co-first authors

*corresponding author

Abstract

Telecom fraud Financial fraud is committed by telephone, Internet and SMS, and it is important to identify and prevent these activities. In this paper, a telecom fraud prediction model is established based on the collected telecom fraud data set. Firstly, the environment in which fraud occurs is analyzed through data visualization technology, and then the influence of bank card usage on fraud probability is discussed by using logistic regression model, and it is found that there is a significant correlation between improper use of bank card and fraud risk. Then, the relationship between transaction pattern and fraud is revealed through the correlation analysis of variables, and the complex relationship between different transaction types and fraud probability is deeply discussed. Finally, convolutional neural network and short-duration memory network model combined with Bahdanau attention mechanism were used to improve the prediction accuracy, and the accuracy of the model was up to 99%. This study not only improves the prediction ability of the model, but also shows the extensive application potential of the model, which provides important technical support and theoretical basis for the identification and prevention of telecom fraud.

Keywords: Telecommunications fraud, Correlation analysis, LSTM, CNN, Attention Mechanism, Model Performance evaluation

1. Introduction

Telecom fraud, especially bank card fraud through telephone, Internet and SMS channels, has become a major problem worldwide. This form of fraud not only poses a direct threat to the financial security of individuals, but also challenges the stability of the financial system. With the rapid development of science and technology, fraud methods are constantly evolving, making traditional prevention methods gradually ineffective. Therefore, the development of efficient forecasting tools to predict and block these fraudulent activities has become an urgent problem.

In recent years, machine learning has shown its powerful data processing and pattern recognition capabilities in many fields, providing new perspectives and methods for solving complex problems. In the field of financial fraud detection in particular, machine learning techniques are able to analyze large amounts of data and identify potential patterns of possible fraud, enabling effective early warning and intervention. Based on this background, this study uses advanced data analysis technology and machine learning methods to explore the detection and prevention strategies of telecom bank card fraud.

2. Logistic regression

2.1. The fundamentals of logistic regression

Logistic regression models use the logistic function (or sigmoid function) to estimate the relationship between a dependent variable and one or more independent variables. The model fits the data through maximum likelihood estimation to find a set of coefficients that maximizes the probability of the observed sample[1].

The form of the logical function is:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{1}$$

Where z is the output of the linear model, usually expressed as:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n \tag{2}$$

The probability P of logistic regression model output is expressed as:

$$P(y=1|X) = \ \sigma(z) = \frac{1}{1+e^{-z}} \tag{3}$$

2.2. Model loss and model training

Logistic regression uses the Log-Loss function to train the model. For binary classification problems, the loss function is expressed as:

$$Log - Loss = -\frac{1}{m}\sum\nolimits_{i=1}^{m} [y^{(i)}\, log\,(\hat{y}^{(i)}) + (1-y^{(i)})\, log\,(1-\hat{y}^{(i)})\,] \tag{1}$$

2.3. The working principle of the model in this paper

In this article, we need to predict whether telecom Fraud occurs based on the following two characteristics: Card: Whether a bank card is used to transfer money on the device, where 1 means yes and 0 means no. Pin: Whether the PIN number of the bank card is used for the transfer transaction, with 1 representing yes and 0 representing no.

The specific formula of logistic regression model is as follows:

$$z = \beta_0 + \beta_1 \cdot Card + \beta_2 \cdot Pin \tag{1}$$

The probability P of predicting telecom fraud is:

$$P(Fraud = 1|Card, Pin) = \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot Card + \beta_2 \cdot Pin)}} \tag{2}$$

3. Model solving

3.1. Solving procedure

According to the model principle of this paper, firstly, data preprocessing is carried out: feature Card and Pin and target variable Fraud are extracted from the data set.

Then the model is trained: the logistic regression model is fitted with the training data set, and the regression coefficients β0,β1,β2 are learned. Then model prediction: The test data set is predicted, the probability of telecom fraud occurring in each sample is calculated and the sample is classified according to the predicted probability (fraud occurs or no fraud occurs). Finally, the model was evaluated by accuracy, accuracy, recall, F1 score and other indicators, and the probability of telecom fraud under different feature combinations was compared to determine which feature combinations were more prone to telecom fraud. [2]

3.2. Result

Python algorithm is designed to achieve the above functions, and the output is as follows:

```
Scenario (Card=0, Pin=0): Fraud Probability = 0.1109
Scenario (Card=0, Pin=1): Fraud Probability = 0.0034
Scenario (Card=1, Pin=0): Fraud Probability = 0.0711
Scenario (Card=1, Pin=1): Fraud Probability = 0.0021
```

3.2.1 Fraud probability

Using a PIN number (Pin=1) significantly reduced the probability of fraud (from 11.09% to 0.34%) in cases where a bank Card was not used for a transfer transaction on the device (Card=0).

In cases where a bank Card was used for a transfer transaction on the device (Card=1), the use of a PIN number (Pin=1) also significantly reduced the probability of fraud (from 7.11% to 0.21%).

3.2.2 Use the validity of the PIN number

When a PIN number (Pin=1) is used, the probability of fraud is significantly reduced[3]. This shows that the use of PIN numbers plays an important role in preventing telecom fraud.

3.2. Model performance

Model accuracy is as follows:

Accuracy=0.9129033333333333

The confusion matrix is shown in Figure 1:

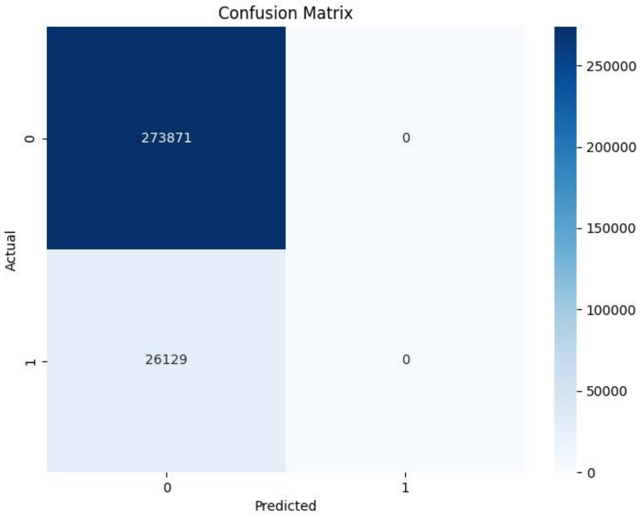


Figure 1. Confusion Matrix of Logistic Regression Model Results

While the model's overall accuracy on the test set was high (91.29%), it failed to correctly predict a single fraud record. This may be due to the low percentage of fraud records in the data set (about 8.74%), causing the model to be more inclined to predict records as non-fraud[4].

The accuracy of fraud prediction using this model is very low, but the classification results have high credibility.

4. Correlation analysis

In order to solve the correlation among variables, Pearson correlation coefficient, Kendall correlation coefficient and Spearman correlation coefficient are commonly used, and their application ranges are shown in Table 1:

Table 1. Scope of application of three correlation coefficients

Correlation coefficient	Scope of application	Data requirement	advantage	shortcoming
Pearson	Continuous data, consistent with normal distribution	Linear relationship, bivariate is continuous type	The calculation is simple and the interpretation is intuitive	Sensitive to outliers and requires normality
Kendall	Ordinal or continuous data does not require a normal	Monotone relation	Strong adaptability to outliers and data distribution	The calculation is more complicated and the efficiency is lower than

	distribution			Pearson's
Spearman	Ordinal data or continuous data that does not satisfy a normal distribution	Monotone relation	It is sensitive to nonlinear relationships and does not require a normal distribution	The calculation is more complicated and the efficiency is lower than Pearson's

The first three fields of the dataset are continuous variables, and the last five fields are categorical variables. For continuous variables, which need to determine whether they are linear and conform to a normal distribution, categorical variables can use a chi-square test[5].

4.1. Linear relationship and normal distribution test

The Shapiro-Wilk test is used to determine whether the data comes from a normal distribution. This test works by calculating the statistics of the sample data and comparing it to the expected normal distribution. If the value p is less than the significance level (usually 0.05), that is, the deviation between the sample data and the normal distribution is significant, then the null hypothesis is rejected and it is considered that the continuous variables Distance1, Distance2, and Ratio do not conform to the normal distribution[6].

The Shapiro-Wilk test statistic W is calculated by the following formula:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1}$$

Shapiro-Wilk test has been able to check whether the three variables have a linear relationship, but in order to ensure the accuracy of the model, K-S test and D'Agostino test are still needed.

The Pearson correlation coefficient is used to measure the correlation between two continuous variables, mainly referring to the linear relationship. It determines the linear correlation between two variables by calculating the ratio of covariance to standard deviation. The Pearson correlation coefficient ranges from -1 to 1, and the closer the absolute value is to 1, the stronger the linear relationship[7].

The Pearson correlation coefficient r is calculated by the following formula:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{2}$$

Python algorithms were written to realize the above tests, and histograms and Q-Q graphs of three continuous variables were drawn, as shown in Figure 2-7.

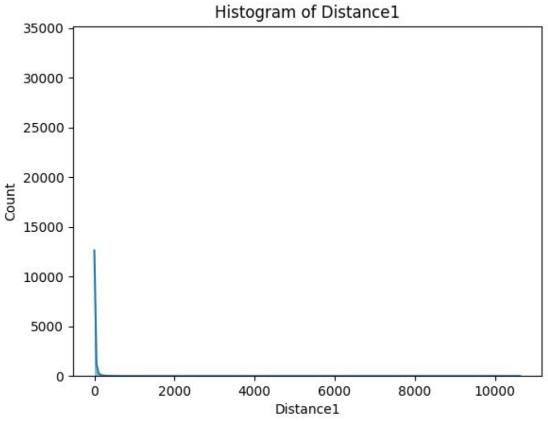


Figure 2. Histogram of Distance1

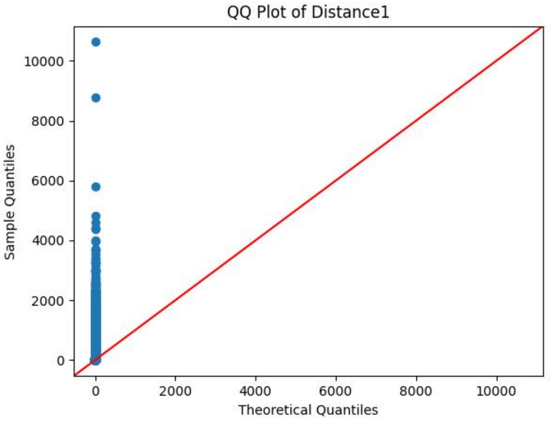


Figure 3. Q-Q plot of Distance1

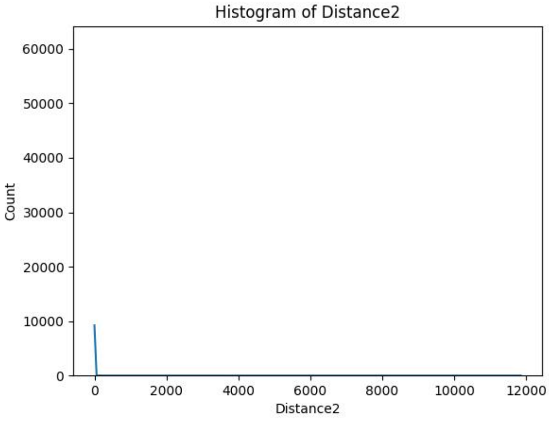


Figure 4. Histogram of Distance2

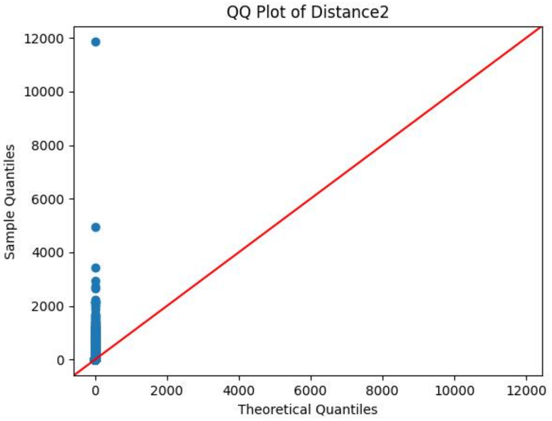


Figure 5. Q-Q plot of Distance2

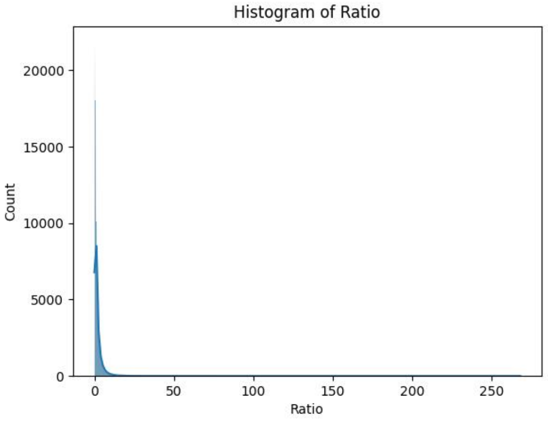


Figure 6. Histogram of Ratio

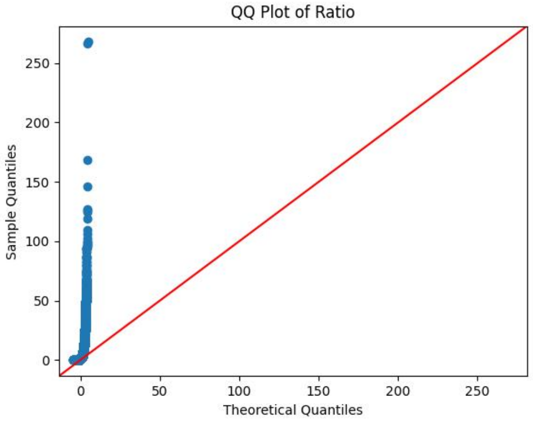


Figure 7. Q-Q plot of Ratio

The Shapiro-Wilk test, K-S test, and D'Agostino test have the following eigenvalues:

Table 2. Characteristic values of three types of tests

	Shapiro-Wilk	K-S	D'Agostino
Distance1	0.0000	0.0000	0.0000
Distance2	0.0000	0.0000	0.0000
Ratio	0.0000	0.0000	0.0000

Draw a histogram to display the data distribution as a series of bars, which can intuitively show the concentration, dispersion, and shape of continuous variables. In theory, a histogram of a normal distribution should present a symmetrical bell shaped curve, with main features including symmetry, unimodal, and tail behavior. Similarly, if the data follows a normal distribution, the straight line connecting the quantiles of the sample data should appear as a 45 ° line on the Q-Q plot[8].

Observing the obtained histograms, the symmetry, unimodality, and tail behavior characteristics of the three continuous variables are not obvious[9]. Observing the Q-Q plot, the distribution lines of the three continuous variables deviate significantly from the 45 ° line.

The eigenvalues of Shapiro Wilk test, K-S test, and D'Agostino test are as follows:

Table 2 shows the eigenvalues of three continuous variables under three different tests, indicating that these variables significantly deviate from normal distribution[10].

Based on the above test results, it can be concluded that Distance1, Distance2, and Ratio do not follow a normal distribution.

The Pearson correlation coefficient obtained is shown in Table 3:

Table 3. Pearson correlation coefficients of three variables

	Distance1	Distance2	Ratio
Pearson correlation	0.1876	0.0919	0.4623

There is almost no linear relationship between Distance2 and Fraud, a weak positive linear relationship between Distance1 and Fraud, and a moderate linear relationship between Ratio and Fraud.

Although the Pearson correlation coefficient shows a certain linear relationship, the results of the normality test indicate that these variables significantly do not follow a normal distribution. Therefore, suitable correlation analysis methods for non normal distributions should be used[11].

4.2. Correlations

4.2.1 Correlation test of continuous variables

Due to the fact that the continuous variable fields in the dataset do not follow a normal distribution and have poor linear relationships, in order to test the correlation between the continuous variable fields and the Fraud field, it is necessary to choose Spearman correlation coefficient tests that are sensitive to non-linear relationships and do not require a normal distribution. Spearman correlation coefficient is a non parametric statistical measure used to evaluate the monotonic relationship between two variables. Spearman correlation coefficient converts raw data into rankings and calculates the correlation between rankings[12].

The Spearman correlation coefficient ϱ is calculated using the following formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{1}$$

Taking the 48th to 53rd data points of Distance1 and Fraud as examples, calculate the Spearman correlation coefficient:

Table 4. raw data

Distance1	Fraud
2.530145	1
21.12612	1
42.73586	0
15.6933	0
43.28131	0

Table 5. Ranking calculation

Index	Distance1	Rank of Distance1	Fraud	Rank of Fraud
1	2.530145	1	1	4
2	21.12612	3	1	4
3	42.73586	4	0	1
4	15.6933	2	0	1
5	43.28131	5	0	1

Table 6. Ranking difference $\boldsymbol{d_i}$ calculation

Index	Rank of Distance1	Rank of Fraud	d_i	d_i^2
1	1	4	-3	9

2	3	4	-1	1
3	4	1	3	9
4	2	1	1	1
5	5	1	4	16

Then,

$$\sum d_i^2 = 9 + 1 + 9 + 1 + 16 = 36 \tag{2}$$

$$\rho = 1 - \frac{6 \cdot 36}{5(5^2 - 1)} = 1 - \frac{216}{120} = 1 - 1.8 =- 0.8 \tag{3}$$

That is, the Spearman correlation coefficient of the selected test set is -0.8.

*This result is only applicable to the selected test set and is used to illustrate the working process of Spearman correlation coefficient test, and does not represent the entire set.

4.2.2 Correlation test of categorical variables

Repeat, Card, and Pin Online fields are categorical variables and should be tested using appropriate methods for categorical variables. Chi squared test is a method used to analyze the correlation between two categorical variables, which determines the independence between variables by calculating the difference between observed frequency and expected frequency. The chi square test is a non parametric test that does not require data to follow a normal distribution or other specific distributions. This makes the chi square test more flexible in processing actual data, and the test method is more reliable and suitable for handling datasets with large sample sizes[13].

The chi square test statistic χ^2 is calculated using the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{1}$$

Construct a contingency table based on the different values of two categorical variables, which displays the observation frequency of each combination.

The calculation formula for the expected frequency E_i is:

$$E_i = \frac{(R_i \times C_j)}{N} \tag{2}$$

Among them, R_i is the total frequency of the i-th row, C_j is the total frequency of the j-th column, and N is the total number of observed values.

Calculate the contribution value of each cell's χ^2 according to the formula, and sum it up to obtain the total χ^2 statistic.

The calculation formula for the degree of freedom df is:

$$df = (r - 1) \times (c - 1) \tag{3}$$

Among them, r is the number of rows and c is the number of columns.

Based on the chi square statistic and degrees of freedom, search for the chi square distribution table or use statistical software to determine the p-value. If the p-value is less than the significance level (usually 0.05), the null hypothesis is rejected that the field is not significantly correlated with the Fraud field.

4.3. Model solution

Build a Python model to solve the above problem, and the results are as follows:

4.3.1.Correlation analysis between continuous variables and telecommunications fraud

The Spearman correlation coefficient between Distance1 and Fraud is 0.0950, indicating a very weak positive correlation between the distance between the transaction location and home and telecom fraud.

The Spearman correlation coefficient between Distance2 and Fraud is 0.0347, indicating a weaker correlation between the distance from the last transaction and telecom fraud.

The Spearman correlation coefficient between Ratio and Fraud is 0.3428, indicating a moderate positive correlation between the current transaction amount and the previous transaction amount and telecom fraud.

The p-values of all these correlation coefficients are 0.0000, indicating significant correlation.

4.3.2.Categorical variables and cardholder test results of telecommunications fraud

The chi square value for Repeat and Fraud is 1.8278, with a p-value of 0.1764, indicating that there is no significant correlation between conducting transactions at the same bank and telecom fraud.

The chi square value of Card and Fraud is 3717.4490, with a p-value of 0.0000, indicating a significant correlation between using bank cards for transactions on devices and telecom fraud.

The chi square value of Pin and Fraud is 10057.4125, with a p-value of 0.0000, indicating a significant correlation between the use of PIN codes and telecom fraud.

The chi square value of Online and Fraud is 36852.0237, with a p-value of 0.0000, indicating a significant correlation between online transactions and telecom fraud.

4.4. Conclusion

The correlation of key variables is as follows:

The card side test results of Repeat and Fraud show that there is no significant correlation between conducting transactions in the same bank and telecom fraud (p-value greater than 0.05).

The chi square test results of Online and Fraud show a significant correlation between online transactions and telecom fraud (p-value far less than 0.05).

The correlation test shows that the five indicators of "transaction amount ratio", "whether transactions are conducted in the same bank", "whether transactions are conducted on devices using bank cards", "whether PIN codes are used", and "whether transactions are conducted online" have a strong correlation with whether telecommunications fraud occurs.

These results indicate that when formulating measures to prevent telecommunications fraud, emphasis should be placed on the transaction amount ratio, whether a bank card is used for transactions on the device (Card), whether a PIN code is used, and whether transactions are conducted online (Online), as they have a strong correlation with telecommunications fraud. The impact of whether transactions are conducted in the same bank (Repeat) on whether telecommunications fraud occurs can be ignored.

5. Establishment of Telecom Fraud Prediction Model

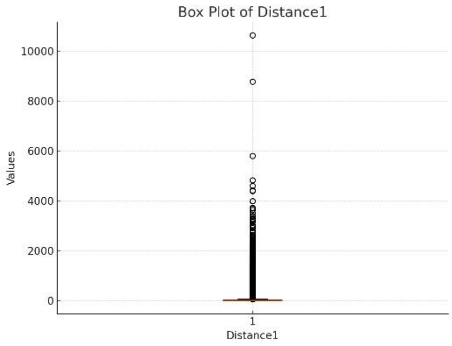
5.1. Data preprocessing

5.1.1 Outlier test

Box plot is a graphical representation method used to display the distribution of a set of data. It provides a visual overview of data distribution by succinctly summarizing the median, quartiles, and outliers of the data.

The core point of using the box plot method to find outliers is to calculate the quartiles, and consider values below Q1-n × IQR or above Q3+n × IQR as outliers, where the standard value of n is 1.5.

Write Python code to draw a box plot of the first seven fields as shown in the following figure:



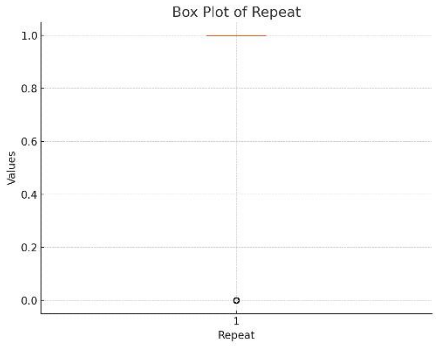
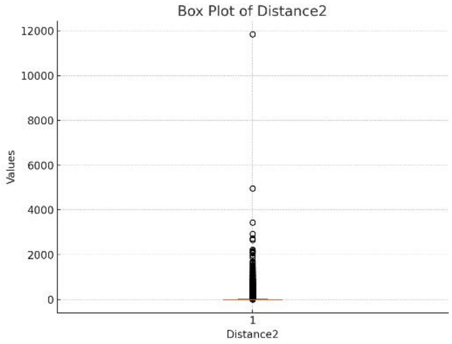


Figure 8. Distance1 Box Line Diagram

Figure 9. Repeat Box Line Diagram



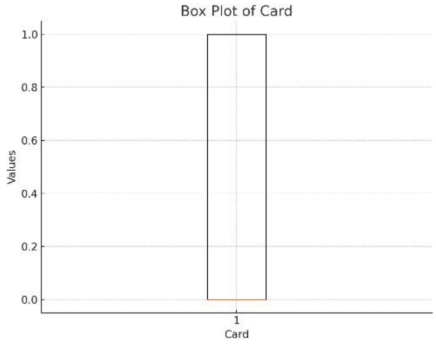


Figure 10. DistanceBox Line Diagram

Figure 11. CardBox Line Diagram

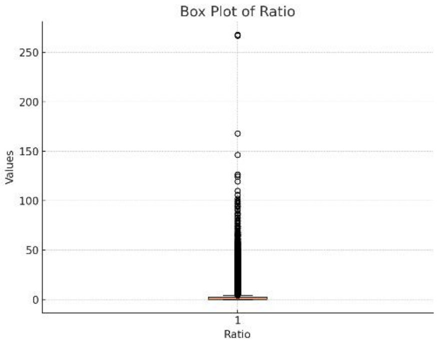


Figure 12. RatioBox Line Diagram

Taking n as the standard value of 1.5, the number of outliers for the first seven variables obtained through programming calculation is shown in Table 7:

Table 7. The number of outliers for the first seven variables

Distance1	103631
Distance2	124367

Ratio	84386
Repeat	118464
Card	0
Pin	100608
Online	0

5.1.2 KNN

K-Nearest Neighbors (k-NN) is a non parametric method that performs classification, regression, or padding operations based on the similarity between neighboring points. When dealing with missing or outlier values, k-NN imputation involves finding the k nearest neighbors of a data point (i.e. the k points closest to that point in the feature space), and then using the values of these neighbors to estimate or replace the missing or outlier values.

In the practical application of this article, KNNImputer in Python is used to achieve fast computation, which automatically finds the k nearest neighbors of points with outliers in each feature. If a feature value is marked as abnormal by the box plot method, it can be adjusted by considering the feature values of its nearest neighbors.

5.2. CNN-LTSM Fusion Prediction Model Based on Attention Mechanism

We have created a CNN-LTSM prediction model based on attention mechanism, with the following model structure:

Input layer: Receive transaction data, whose shape depends on specific time steps and feature numbers.

CNN layer: The first layer is a convolutional layer that uses ReLU activation function to capture local features in transaction data.

LSTM layer: receives the output of the CNN layer, which helps the model understand the temporal dynamics in the data.

Attention layer: a custom Bahdanau attention layer that enhances the model's focus on important time steps by calculating context vectors and attention weights.

Output layer: A fully connected layer using sigmoid activation function to predict whether a transaction is fraudulent.

Training and Optimization

The model is trained using the Adam optimizer with a loss function of binary cross entropy, which is suitable for binary classification problems. Batch normalization and Dropout techniques were used during the training process to prevent overfitting and ensure that the model has good generalization ability even on unseen data.

5.2.1 CNN

Use CNN (Convolutional Neural Network) to process input transaction data. CNN is a powerful neural network architecture commonly used in image processing and any form of multidimensional array data processing, including time series and transaction data. Its main purpose is to extract local features from data through convolution operations.

The CNN layer mainly consists of the following components:

Convolutional layer: Using a set of learnable filters (also known as convolution kernels) to perform convolution operations directly on the raw input data, these filters can capture local features of the data. Each filter slides over the input data one by one, calculates the dot product of the filter and the data, and generates an output called a feature map.

Activation function: Nonlinear activation functions, such as ReLU (Rectified Linear Unit), are usually applied after convolution to increase the network's nonlinear ability and enable it to learn more complex features.

Pooling layer: In some CNN architectures, pooling layers (such as max pooling) are used to further reduce the spatial dimension of feature maps and enhance the robustness of the model to small positional changes.

In this prediction model, the use of CNN layers can help automatically identify complex patterns and features that may indicate fraudulent behavior, without the need to manually design feature extraction rules. The model first receives transaction data as input, then automatically learns features that help identify fraudulent behavior from these transaction data through its convolutional kernel, and finally outputs the feature map to the next layer.

5.2.2 LSTM

In the field of deep learning, Long Short Term Memory (LSTM) is a special type of Recurrent Neural Network (RNN) that is particularly suitable for processing and predicting important events with long time gaps in sequence data. The main advantage of LSTM is its ability to learn long-term dependencies, solving the gradient vanishing problem that traditional RNNs may encounter during long sequence training.

Long Short Term Memory Network (LSTM) is a special type of Recurrent Neural Network (RNN) primarily used for tasks involving processing and predicting sequential data, with the aim of capturing time series dependencies in transaction data.

The core components of LSTM include:

The LSTM unit consists of several key components that work together to allow the network to maintain long-term internal states while processing data:

Forget Gate: Determine which information should be discarded from the cellular state. Controlled by a Sigmoid neural network layer, it looks at the previous hidden state and current input, outputting a value between 0 and 1, indicating how much old information is retained in each cell state.

Input Gate: Update cell status. Firstly, a Sigmoid layer determines which values will be updated, and secondly, a Tanh layer creates a new candidate value vector, which will be added to the state.

Cell State: It is the "memory" part of the network that runs through the entire chain with only slight linear interactions, allowing information to flow through the network without much change.

Output Gate: The part that determines the output based on the cell state. The output is processed by tanh (compressing values between -1 and 1) based on the cell state, and then passed through a Sigmoid gate to determine which part of the information will be output.

As shown in Figure 13.

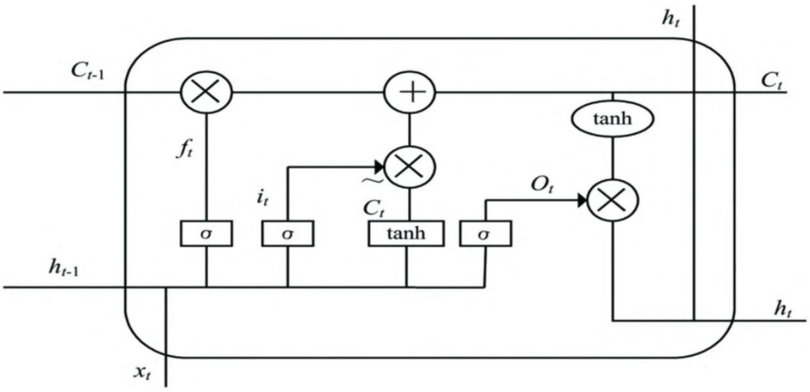


Figure 13. Long short-term cellular memory structure

In this prediction model, the LSTM layer is responsible for capturing time series dependencies in transaction data. The LSTM layer receives features output from the CNN layer, which represent the spatial properties of transaction data at each time step. By maintaining the internal state (cell state), the LSTM can remember and utilize past information to help the model identify fraudulent behavior patterns that may span multiple time steps. Its output provides a comprehensive feature representation that combines spatial features and temporal dynamics for each time step.

5.2.3 Introduce custom attention mechanism

The Bahdanau attention mechanism improves the processing ability of neural network models on sequential data by paying varying degrees of attention to different parts of the input. The core is to create a layer that can learn how to allocate attention weights, which can be achieved through the following steps:

Weight Matrix: Define three weight matrices (for queries (hidden state), keys (encoder output), and values) that will be used to generate attention scores.

Scoring function: uses the weighted sum of queries and keys to calculate scores, implemented through a small neural network or single-layer perceptron.

Softmax layer: Apply the softmax function to the score and convert it into probability distributions that represent the importance of each input.

Context vector: Use the generated probability distribution (attention weights) to create a weighted sum, which will serve as the final output and be provided to the next layer of the model.

This custom attention mechanism will allow the model to focus more on the most critical parts of the input data for the prediction task.

5.2.4 Add techniques to prevent overfitting

(1) Regularization

Dropout is a common regularization technique that prevents neural networks from overfitting by randomly "dropping" (i.e. setting to zero) a portion of the activation values of neurons during the training process. This method forces the network to learn more robust features because it cannot rely on any one neuron, as neurons may be randomly removed during training.

(2) Batch normalization

Batch Normalization is another technique that accelerates the training process and improves the stability of the model by normalizing the input of the normalization layer. It is achieved by reducing the offset of internal covariates, applied before the activation function of each layer.

5.2.4 Performance evaluation of the model

The training process is set to perform ten complete iterations on the entire training set, during which the model will see the training data multiple times and have the opportunity to adjust its weights and biases each time to reduce prediction errors.

The process of each epoch includes:

Forward propagation: At this stage, input data flows through the model to generate output.

Calculate loss: The difference between the predicted output of the model and the actual value is calculated through a loss function. This loss represents the current performance of the model.

Backpropagation: By using a loss function to calculate gradients for each weight, and then using these gradients to update the model's weights, this is accomplished through optimization algorithms such as SGD, Adam, etc.

Table 8. Model performance of ten Epoch processes

	Training accuracy	training loss	validation accuracy	validation loss
Epoch 1/10	0.9732	0.0701	0.9921	0.0205

Epoch 2/10	0.9932	0.0176	0.9957	0.0139
Epoch 3/10	0.9948	0.0135	0.9969	0.0083
Epoch 4/10	0.9956	0.0111	0.9974	0.0078
Epoch 5/10	0.9961	0.01	0.9962	0.0085
Epoch 6/10	0.9964	0.0091	0.9974	0.0067
Epoch 7/10	0.9967	0.0084	0.9966	0.0082
Epoch 8/10	0.9970	0.0076	0.9938	0.0162
Epoch 9/10	0.9970	0.0077	0.998	0.0053
Epoch 10/10	0.9971	0.0072	0.9985	0.0047

The training results show that both the training and validation accuracy are high, and the validation loss gradually decreases, indicating that the model gradually converges during the training process and has good generalization ability. Starting from the 5th epoch, the validation accuracy has been very close to 100%, and the validation loss is also very low, indicating that the model can learn patterns in the data very well.

Accuracy: 0.9984
Precision: 0.9933
Recall: 0.9888
F1 Score: 0.9911

In addition, the confusion matrix is used to intuitively understand the prediction accuracy. The confusion matrix displays the performance of the model in each category, including true positives, false positives, true negatives, and false negatives. By using the confusion matrix, we can intuitively see which categories the model performs well in and which categories have problems. Then draw the ROC curve, which is an important tool for evaluating the classification performance of the model, especially when the dataset is imbalanced, along with the receiver operating characteristic curve (ROC curve) and AUC (area under the curve). The ROC curve is plotted by calculating the true positive rate (recall) and false positive rate (false positive rate) at different thresholds. The AUC value provides the ability of the model to distinguish between positive and negative classes, and the higher the AUC value, the better the performance of the model. The calculation formula for AUC is the area under the ROC curve.

As shown in the following figure:

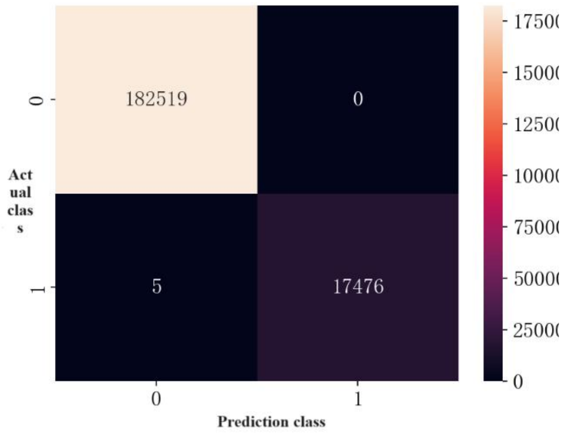


Figure 14. Confusion matrix diagram

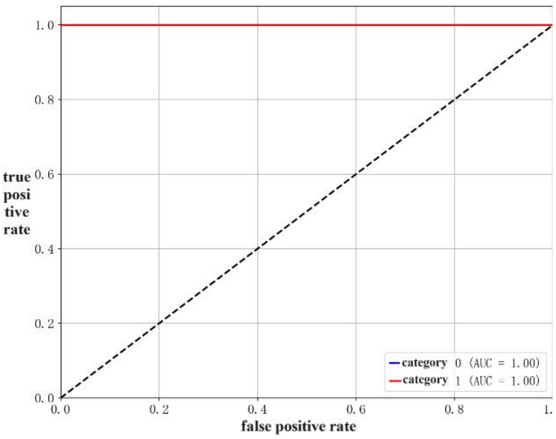


Figure 15. ROC curve chart

From the confusion matrix, it can be seen that the model performs very well in avoiding misclassifying non fraud categories as fraud (false positives are 0), which means its accuracy is very high. However, 5 fraud cases were mistakenly classified as non fraud, indicating that although the sensitivity (recall) of the model is also high, there is still room for improvement. The AUC value is approximately 1, indicating that the model has strong classification ability and can distinguish fraud and non fraud cases with extremely high accuracy regardless of the threshold setting.

This performance indicates that the model is almost perfect for the current dataset, but its robustness still needs to be validated on new or more diverse data. The performance of the model may be due to the highly representative training data or to some extent simplifying the complexity of the real world. Perfect performance may sometimes mean overfitting, especially if the training data differs from the data in the actual application scenario.

6. Conclusion

This study successfully constructed a high-precision telecommunications fraud prediction model by integrating multiple machine learning techniques. We first used data visualization technology to deeply analyze the environment and patterns of telecommunications fraud, and then explored the relationship between bank card usage and fraud probability through logistic regression models. In addition, through in-depth analysis of variable correlation, multiple key factors such as the close relationship between transaction patterns and telecommunications fraud have been revealed.

In the construction of the model, we used convolutional neural networks and long short-term memory network models, and introduced the Bahdanau attention mechanism to optimize the model's processing ability and prediction accuracy for time series data. The application of this ensemble learning method significantly improves the performance of the model, achieving a prediction accuracy of 99%.

The research results indicate that the application of modern machine learning technology can effectively predict and prevent telecommunications fraud, providing strong technical support for banks, government departments, and security agencies. In addition, our research provides new perspectives and methods for future research in this field, demonstrating the enormous potential of data analysis and machine learning techniques in solving complex problems.

References

[1] Chen Xiaolei, Xu Hui Construction of Evidence System for Telecommunications Fraud Cases [J] Network Security Technology and Applications, 2024 (5): 139-141.

[2] Qin Yutian The logical development and criminal control of telecom network fraud with fraud as the core charge [J/OL] Journal of Taiyuan University (Social Sciences Edition), 2024, 25 (4): 95-104.

[3] Zhang Hao Analysis of Ecological Governance of the "Black and Grey Industry Chain" in Telecommunications Network Fraud Crimes [J] Network Security Technology and Applications, 2024 (4): 148-150.

[4] Xiao Song, Huang Jiewu First order approximate knife cut correction ridge estimation in binary logistic regression models [J] Journal of Hunan University of Arts and Sciences (Natural Science Edition), 2024, 36 (2): 6-13.

[5] Zhang Jian, Liu Lin, Li Qian Research on Optimization of M-Score Model Based on Nonlinear Logistic Regression [J] Business Accounting, 2024 (10): 83-86.

[6] Liao Wen Research and application of Internet financial user churn prediction model based on data mining [D/OL] South China University of Technology, 2021.

[7] Zhao Zemin Research on Smart Habitat Mode and Behavior Prediction Based on Data Mining Technology [D/OL] Harbin Institute of Technology, 2021.

[8] Li Hao, Zhao Qing, Cui Chenzhou, etc Stellar Spectral Classification Algorithm Based on CNN and LSTM Composite Deep Model [J] Spectroscopy and Spectral Analysis, 2024, 44 (6): 1668-1675.

[9] Zhao Zemin Research on Smart Habitat Mode and Behavior Prediction Based on Data Mining Technology [D]. Harbin Institute of Technology, 2021.

[10] Li Huifeng, Li Tiecheng, Li Weixun. Application of Fuzzy Theory in the Evaluation of Communication Networks in Intelligent Substations [J]. Mechanical Design and Manufacturing, 2024 (04): 28-32+37.

[11] Cheng Junhan, Wang Shuli, Cai Zhiyuan. Remaining service life prediction of lithium batteries based on AE-LSTM [J]. Electrical and Energy Efficiency Management Technology, 2023 (09): 69-75.

[12] Liu Shengjiu, Liang Shupeng, Liu Ying, etc Property Analysis and Application Research of Hypergraph Entropy [C]//Chinese Society of Automation. Proceedings of the 2023 China Automation Conference [Publisher unknown], 2023: 6.

[13] Zhang Haiyang, Chen Yuming, Zeng Nianfeng, et al. Credit Card Fraud Detection Based on XGBoost and LR Fusion Model [J]. Journal of Chongqing University of Technology (Natural Sciences), 2024, 38 (03): 195-200.