Paper Type: Original Article

Exploring Machine Learning and Data Analysis Strategies for Telecom and Bank Card Fraud Patterns.

Jiangyang Xu[1#], Shuying Jie[2#], Runda Xiong[3#]

1.Changzhou Institute of Technology, School of Computer Science and Information Engineering, Changzhou, Jiangsu

2.School of Physical Electronics and Information, Gannan Normal University, Ganzhou,Jiangxi Province

3.Jinshan College, Fujian Agriculture and Forestry University, Fuzhou City, Fujian Province

#Co-first author

Abstract

Telecom fraud is a criminal activity that utilizes phones, the internet, and text messages to deceive victims into transferring funds. With the advancement of technology, telecom fraud methods continue to evolve, making the identification and prevention of these fraudulent activities increasingly important. This paper combines data provided by the organizing committee for data analysis and the establishment of a telecom and bank card fraud prediction model.

In Problem 1, data visualization techniques were used to analyze the environment in which telecom and bank card fraud occurs. Pie charts and bar charts were created to display the frequency of fraud occurrences and the ratio of online to offline fraud. The analysis revealed a higher incidence of telecom fraud in online environments, providing a basis for subsequent preventive strategies.

In Problem 2, we used a logistic regression model to analyze the impact of bank card usage (whether the card was used for transfers on a device and whether a PIN code was used for transactions) on the probability of telecom fraud. The results showed that using a PIN code significantly reduced the likelihood of fraud, suggesting that strengthening the use of PIN codes is an effective strategy in preventing telecom fraud.

In Problem 3, we tested the linearity and normality of variables, using Spearman's correlation analysis and chi-square tests to reveal the correlation between factors such as the transaction amount ratio, whether transactions were conducted with the same bank, and whether transactions were made online. The in-depth analysis of these indicators aids in more accurately identifying potential fraudulent activities and provides feature importance for the prediction model.

In Problem 4, convolutional neural network (CNN) models, long short-term memory (LSTM) models, and attention mechanisms were introduced. We employed stacking ensemble learning to integrate models, improved the attention mechanism, and added measures to prevent overfitting. Finally, we calculated accuracy, precision, recall, and F1 scores for each model. The results showed that the telecom and bank card fraud prediction model established in this paper achieved an accuracy of 99.99%, effectively predicting fraud. Based on the data analysis throughout this paper, we provide recommendations for public security departments, banks, and citizens to reduce the probability of telecom fraud.

In conclusion, this paper applies modern data analysis and machine learning techniques, with particular innovations in model integration and the application of attention mechanisms. Not only does this improve the prediction ability of the models, but it also validates their applicability and

robustness through accuracy and error analysis. The research results are compared with existing technologies in a horizontal comparison and demonstrate the model's potential for future expansion, providing a promising outlook for telecom and bank card fraud detection.

Keywords: Telecom fraud, correlation analysis, neural networks, machine learning, model performance evaluation.

Introduction

Background

Telecom fraud is an elaborate fraud against targeted victims using modern communication technology, such as telephone, Internet and text messages. The main feature of this form of crime is the [1] through remote non-contact means, and the fraudsters do not need to face to face with the victim, thus greatly reducing the risk of being directly identified and tracked.scammers often disguise themselves as individuals or institutions with higher credibility, such as government officials, legal agencies, well-known companies or bank employees, to win the trust of victims [2]. With the development of technology, the means of telecom fraud are also constantly changing, and new fraud methods such as false network platforms and APP fraud are increasing. Fraudsters may use social media, fake official websites or lure victims by sending links with malware.

Figure 1. Telecom fraud

In response to this problem, governments and relevant departments have taken a variety of measures, including strengthening public risk awareness education, improving the safety protection measures of financial institutions and promoting the improvement of laws and regulations. In addition, transnational cooperation is also intensifying to track and combat transnational telecom fraud criminal networks. Nevertheless, telecom fraud is still a complex and continuously evolving challenge that requires the concerted efforts and continuous attention of [3] from all sectors of society.

Restatement of the problem

Problem 1, data visualization. Make a fan chart to clearly show the proportion of occurrence and no telecom card fraud in the data provided, and create a bar chart to compare the frequency of telecom fraud in online and offline environments, so as to reveal which environment is more vulnerable.

Question 2, correlation strength analysis. Through the in-depth analysis of the data, we determine the two cases of bank card, which is more likely to encounter fraud, and discuss whether the PIN number of bank card can reduce the probability of fraud.

Question 3, correlation analysis. In all the recorded telecom fraud, which indicators have a strong correlation with the possibility of telecom fraud. It is necessary to evaluate whether bank card transfer transactions occur in the same bank and whether they are online transactions are associated with telecom fraud.

Problem four, fraud prediction model construction. Based on all available index data, build a prediction model of telecom bank card fraud. Appropriate indicators were selected as input to the model to accurately segment the training and test sets. After establishing the model, its accuracy in predicting telecom fraud is calculated, and the effectiveness of the model is evaluated.

Problem analysis

Analysis of question 1

Problem 1 The main task is data visualization. Use the algorithm to count the total number of telecom fraud and no telecom fraud and calculate the proportion of fraud, record the specific number of online and offline fraud records, and print out these data for inspection. Next, the statistical values are visualized to draw a fan chart of the proportion of "whether there is telecom bank card fraud", as well as the bar chart of the number of "online" and "offline" fraud in the cases of telecom fraud.

Analysis of question two

The second main task is to analyze the strength of the correlation. First, relevant features (Card and Pin) and target variables (Fraud) need to be extracted from the data set and the proportion of fraud occurs in different situations calculated. We grouped the data according to whether to use the bank card to transfer transactions on the device and whether to use the PIN number to transfer transactions, and counted the proportion of telecom fraud in each group. Next, by calculating the probability of fraud in different groups, we can directly compare the risk of telecom fraud in different situations, and judge whether using the PIN number can reduce the probability of being cheated. Through these analyses, we can understand which cases are more prone to telecom fraud, and evaluate the effectiveness of using PIN numbers in reducing telecom fraud.

Analysis of question three

The three main task is correlation analysis. To check whether the continuous variable meets a normal distribution and has a linear relationship with the target variable Fraud to select a suitable correlation analysis method. For continuous variables, the appropriate correlation analysis method was selected based on the results of normality and linear tests; for categorical variables, the chi-square test was used to assess the association with the target variable Fraud.

Continuous variables were tested for normality using the Shapiro-Wilk test, the Kolmogorov-Smirnov test, and the D 'Agostino's K-squared test. If the p-value is less than the significance level (usually 0.05), the null hypothesis is rejected and the data does not fit a normal distribution. The Pearson's correlation was used to measure the linear relationship between the continuous and target variables using Pearson's correlation coefficient (Pearson Correlation Coefficient). The Pearson correlation coefficient ranges from-1 to 1, and the closer the value is to $\pm$ 1, the stronger the linear relationship is. If the continuous variable does not fit a normal distribution or has a nonlinear relationship, Spearman's correlation coefficient (Spearman Correlation Coefficient) was used.

Analysis of question four

The main task of question 4 is to create predictive models. To establish a "telecom bank card fraud prediction model", calculation accuracy, and combined with our data analysis results to provide opinions to the public security department, banks and citizens three parties. First do data cleaning, screen for outliers and fill. Multiple models (CNN, LSTM) are selected for learning, and then fusion algorithms are used to integrate the models. The accuracy and robustness of prediction are improved through integrated learning methods, and multiple indicators such as precision, precision, recall rate and F1 score are evaluated to ensure its effect and practicability. Finally, the model is explained, so as to give the public security department, banks and citizens to put forward our suggestions and reduce the probability of telecom fraud.

Model preparation

Model hypothesis

To simplify the given problem and modify it to be more suitable to simulate real-life conditions, we make the following basic assumptions.

Suppose that even in fraudulent transactions, abnormal behavior (e. g., very large transaction amount or atypical transaction frequency) follows some known statistical distribution (e. g., Gaussian distribution). This helps the model to use statistical methods to identify the boundaries between standard behavior and abnormal behavior.

Suppose that the model parameters (e. g., weight, bias) remain stable throughout the prediction period of the model and do not need to be adjusted frequently due to the introduction of new data. This is often achieved with appropriate model calibration and update strategies in practice.

The chosen loss function is assumed to reasonably quantify the prediction error and the model performance. For example, the logarithmic loss function is assumed to effectively handle the probability estimation error of the model output.

Assuming that the dataset used to train the model can statistically represent the characteristics of the entire target population, that is, the samples used for the model training are randomly and fairly selected from the whole data generation process.

Symbol description

The key mathematical symbols used in this paper are listed in Table 1.

Table 1 shows the symbol description

symbol	explain
Distance1	The distance of the bank card transfer transaction place from the home
Distance2	The distance from the last bank card transfer transaction that occurred
Ratio	The ratio of the amount of the bank card transfer transaction to the amount of the last bank card transfer transaction

Repeat	Whether the bank card transfer transaction occurs in the same bank, 1 means yes and 0 means no
Card	Whether to use the bank card to transfer transactions on the device, 1 means yes, 0 means no
Pi	Whether to use the PIN number of the bank card for transfer, 1 means yes and 0 means no
Online	Whether it is an online bank card transfer transaction, 1 represents yes, and 0 means no
Fraud	Whether this bank card transfer is a telecom fraud, 1 represents yes, 0 represents no
N_{fraud}	The number of records with fraud (i. e. Flood field value of 1) in the data set
N_{no_fraud}	Number of records without fraud (i. e. Flood field value of 0)
N_{online}	The number of records online (i. e. Online 1) in the data set
$N_{offline}$	The number of records in the offline (i. e. Online is 0) in the data set

* Other symbol instructions will be given in the text.

Solve the problem

Data set analysis

Basic information of the data

The data set contains 1,000,000 records of transactions using telecom bank cards, with a total of 87,403 without telecom fraud and 912,597 without telecom fraud.

Data preprocessing using the pandas library yields the following information:

Table 2. Basic information of the data set

Column	Non-Null Count	Dtype
Distance1	1000000 non-null	float64
Distance2	1000000 non-null	float64
Ratio	1000000 non-null	float64
Repeat	1000000 non-null	int64
Card	1000000 non-null	int64
Pi	1000000 non-null	int64

Online	1000000 non-null	int64
Fraud	1000000 non-null	int64

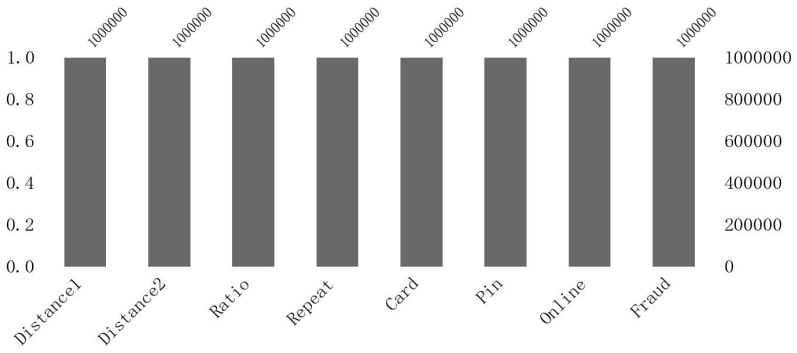

Figure 2 Missing values for viewing

The data set structure is complete with no missing values, the data type (Dtype) includes floating point number (float64) and integer (int64), the first three fields are continuous variables, and the last four fields are categorical variables as expected.

Description of the statistics of the data

Table 3 The Data set describes the statistics

field	Distance1	Distance2	Ratio	Repeat	Card	Pin	Online	Fraud
count	1000000	1000000	1000000	1000000	1000000	1000000	1000000	1000000
mean	26.62879	5.036519	1.061098	0.650552	0.650552	0.500486	0.650552	0.087403
std	65.39078	25.84309	1.272125	0.476796	0.476796	0.5	0.476796	0.282425
min	0.004874	0.000118	0.000132	0	0	0	0	0
25%	3.878008	0.296671	0.230695	0	0	0	0	0
50%	9.96776	0.99865	0.606584	1	1	1	1	0
75%	25.74399	3.355748	1.322953	1	1	1	1	0
max	10632.72	11851.1	179.342	1	1	1	1	1

The means of Distance1 and Distance2 are 26.63 and 5.04, respectively, with large standard deviations, indicating a wide distribution of these distance values.

The average Ratio is 1.06, indicating that the ratio of the current transfer amount to the last amount is slightly greater than 1.

The mean values of Repeat, Card, Pin, and Online are 0.65,0.65, and 0.87, respectively, indicating that most transactions occur online, and most of the transactions use cards and PIN.

The average mean is 0.0874, indicating that about 8.74% of recorded transactions are telecom scams.

Model building

Solve the proportion of telecom bank card fraud

Data set Card Fraud contains n records, that is, n telecom bank card transaction related data, each record contains the Fred field, this field indicates whether the telecom bank card fraud has occurred, the value of 1 means the occurrence, and 0 means that it has not occurred.

The proportion of telecom bank card fraud Occuring Proportion is:

$$\text{Occuring Proportion} = \frac{N_{fraud}}{N} \tag{1}$$

among,

$$N = N_{fraud} + N_{no\_fraud} \tag{2}$$

The number of telecom bank card fraud occurring "online" and "offline"

Online Whether the field is an online transfer transaction record? For the data with the Flood field value of 1 in the data set, Online field indicates that the telecom bank card fraud occurs online or offline, Online=1 means online, and Online=0 means offline.

Then,

$$N_{fraud} = N_{online} + N_{offline} \tag{1}$$

Algorithm establishment

Use the value_counts function in the pandas library to count the number of 0 and 1 in the FDA field and fan chart showing the ratio of fraud and non-fraud, filter out the data with the FDA field value of 1, that is, cases of telecom fraud, count the number of 0 and 1 in the Online field, and draw a bar chart to show the number of online and offline fraud

Model solution

Computing result

The output result of the model is shown in Table 4:

Table 4 shows the results of the classification statistics in question 1

project	statistic
Total fraud cases	87403
Total non-fraud cases	912597
Online fraud cases	82711

Offline fraud cases	4692

Namely, among these telecom bank card transfer records, a total of 87,403 telecom bank card fraud occurred, of which 82,711 occurred online and 4,692 occurred offline.

Then,

$$\text{Occuring Proportion} = \frac{87403}{1{,}000{,}000} = 8.7403\% \quad (1)$$

$$N_{online} = 82711 \quad (1)$$

$$N_{offline} = 4692 \quad (3)$$

Visualization display

The fan chart of the ratio of "with or without telecom bank card fraud" is shown in Figure 3, in the case of telecom bank card fraud, the bar chart of the number of telecom fraud occurring on "online" and "offline" is shown in Figure 4:

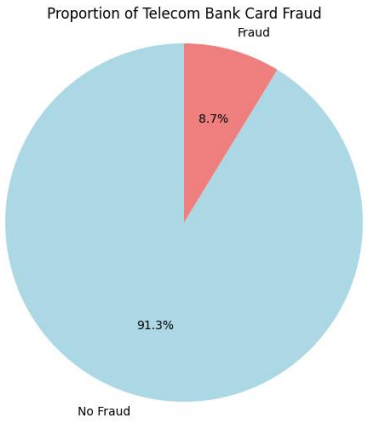

Figure 3 is a fan diagram of the proportion of telecom bank card fraud

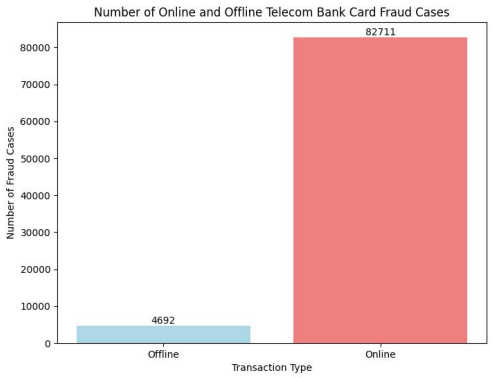

Figure 4 Bar chart of the online and offline number of telecom fraud occurred

Interpretation of result

There are 1000000 0 records in a given dataset. Among them, the total number of telecom fraud (Fraaud = 1) is 87403, and the total number of no telecom fraud (Fraud = 0) is 912597; In the records of telecom fraud, the number of records online (Online=1) is 82711, and the number of records offline (Online=0) is 4692. Telecom bank card fraud cases accounted for about 8.7%, of which online fraud accounted for about 94.6%.

The overall proportion of telecom fraud is relatively small, but its absolute number is still considerable, and effective measures need to be taken to prevent it. Most telecom bank card scams occur through online transactions, indicating that the security of online transactions is an important issue. Public security departments need to strengthen the monitoring and crackdown on online transactions and carry out targeted publicity and education activities to improve the public's awareness of telecom fraud prevention.

Solution to question two

Model building

Logic regression model

Logistic regression (Logistic Regression), also known as log-odds regression, is a widely used generalized linear regression analysis model, especially suitable for dichotomy problems. Despite having the word "regression" in the name, logistic regression is actually a classification algorithm, [4]. It classifies it by estimating the probability of an event.

Rationale of the logistic regression model

The logistic regression model uses a logistic function (or sigmoid function) to estimate the relationship between the dependent variable (usually a dichotomoutcome) and one or more independent variables (features). The model fits the data by maximum likelihood estimation (MLE) to find a set of coefficients that maximize the probability of the observed sample being [5].

The logical function takes the form of:

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

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

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

among,

Is theβ_0 intercept (intercept)

Is the regression coefficient (c$\beta_1, \beta_2, \beta_3, \ldots, \beta_n$oefficients)

Is a characteristic (features)$x_1, x_2, x_3, \ldots x_n$

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

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

Model loss and model training

The logistic regression was trained using the log loss function (Log-Loss). For binary classification

problems, the loss function is expressed as:

$$\text{Log} - \text{Loss} = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)})] \quad (1)$$

among,

m is the number of samples

Is the try$^{(i)}$ue label of the i-th sample

Is the pred$\hat{y}^{(i)}$icted probability of the i-th sample

How the model of this question works

In this question, we need to predict whether telecom fraud (Fred) occurs, based on the following two characteristics:

Card: whether to use the bank card to transfer transactions on the device, 1 means yes and 0 means no

Pin: whether to use the PIN number of the bank card for transfer transactions, 1 means yes and 0 means no

The specific formula of the logistic regression model is:

$$z = \beta_0 + \beta_1 \cdot \text{Card} + \beta_2 \cdot \text{Pin} \quad (1)$$

The predicted probability of telecom fraud P is:

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

Model solution

Solution procedure

Combining the model principle of this problem:

data preprocessing:

Features Card and Pin and the target variable Fred were extracted from the dataset.

2) Model training:

The training data set was fitted to the logistic regression model, learning $\beta_0, \beta_1, \beta_2$the regression coefficients

3) Model prediction:

A) forecast the test data set and calculate the probability of telecom fraud for each sample.

And b) classify the sample according to the predicted probability (fraud or no fraud).

4) Model evaluation:

A) Model performance is evaluated by accuracy, precision, recall, F1 score and other indicators.

BB) Calculate and compare the probability of telecom fraud under different feature combinations, and judge which feature combinations are more likely to occur telecom fraud.

Bear fruit

Design Python algorithm implements the above functions, and the output results are 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

1) Fraud probability

Without a bank card (Card=0), using the PIN number (Pin = 1) significantly reduced the probability of fraud (from 11.09% to 0.34%).

Using a PIN number (Pin = 1) also significantly reduced the probability of fraud (from 7.11% to 0.21%) (Card=1).

2) Effectiveness of using PIN numbers

With the PIN number (Pin = 1), the probability of fraud decreased significantly. This indicates that the use of PIN numbers plays an important role in preventing telecom fraud.

Model performance

The model accuracy is as follows:

Accuracy=0.9129033333333333

Figure 5:

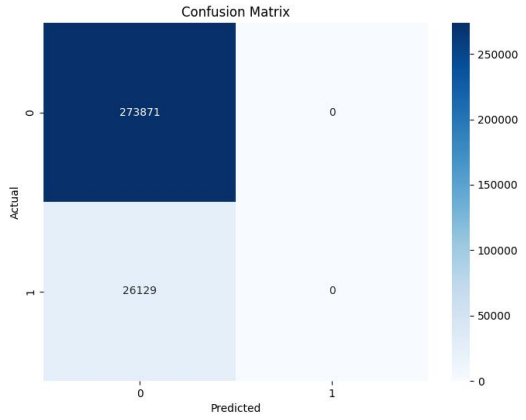

Figure 5 Confusion matrix plot of the logistic regression model results

Although the model had a high overall accuracy on the test set (91.29%), it failed to correctly predict any fraud record. This may be due to the low proportion of fraud records in the data set (about 8.74%), which leads the model to predict the records as non-fraud.

The accuracy of predicting the occurrence of fraud by using this model is very low, but the classification results have high confidence.

Solution to question three

Model building

Problem three mainly solves the correlation of the fields, such as Pearson correlation coefficient, Kendall correlation coefficient and Spearman correlation coefficient [6,7]. The scope of application of the three is as shown in Table 5:

Table 5 Scope of application of the three correlation coefficients

correlation coefficient	scope of application	Data requirements	merit	shortcoming
Pearson	Continuous data, which conform to the normal distribution	Linear relationship, bivariate are continuous	The calculation is simple and intuitive	Sensitive to outliers and requires normality
Kendall	Ordinal data or continuous data, no normal distribution is required	Monotonic relationship	Adapadaptability to outliers and data distribution	The calculation is complicated and has low in relative Pearson efficiency
Spearman	Ordinal data or continuous data that do not satisfy a normal distribution	Monotonic relationship	Sensitive to nonlinear relationships and does not require a normal distribution	Is greatly affected by the outliers, especially if the data volume is small

The first three fields of the data set are continuous variables, and the last five fields are categorical variables. For continuous variables, it is necessary to determine whether they are linear and fit to a normal distribution, and categorical variables can be tested using the chi-square test.

Linear relationship with a normal distribution test

The Shapiro-Wilk test was used to determine whether the data were derived from a normal distribution. The test is performed by calculating the statistic $\bar{p}$ of the sample data and comparing it to the expected normal distribution. If the value is less than the significance level (usually 0.05), that is, the sample data is significantly deviated from the normal distribution, the null hypothesis is rejected and the continuous variables Distance1, Distance2, and Ratio do not meet the normal distribution.

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

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

among,

Is the i-$x_{(i)}$th value after the sample ordering.

For thea_i weight.

Fonr the number of samples.

For $\bar{x}$sample mean.

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

The Pearson's (Pearson) correlation coefficient is used to measure the correlation between two continuous variables, mainly for the linear relationship. It determines the linear correlation between two

variables by calculating the ratio of the covariance to the standard deviation. The Pearson correlation coefficient ranges from-1 to 1, and the closer to 1, the stronger the linear relationship.

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

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

among,

And are the sa$x_i y_i$mple values of the two variables, respectively,

And are $\bar{x}\bar{y}$the means of the two variables, respectively,

And n is the number of samples.

The Python algorithm was written to implement the above test, and to draw the histogram and Q-Q diagram of the three continuous variables as follows:

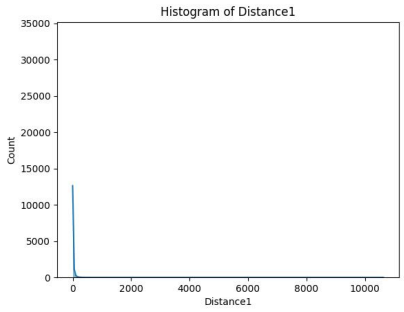

Figure 6 The histogram of Distance1

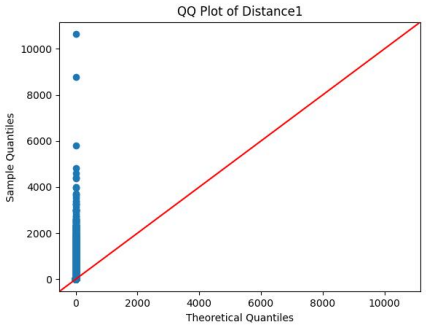

Figure 7 The Q – Q plots of Distance1

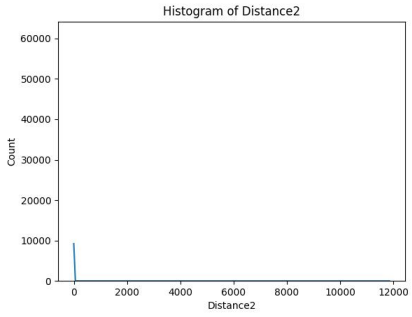

Figure 8 The histogram of Distance2

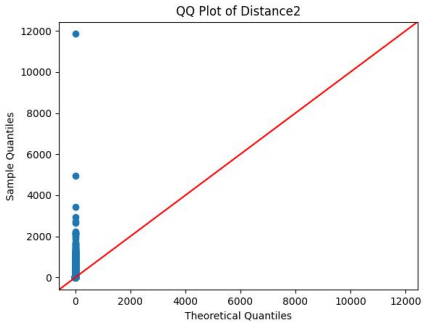

Figure 9 The Q – Q plots of Distance2

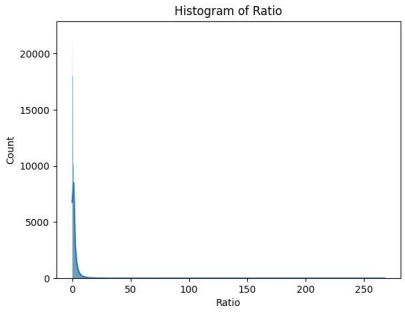

Figure 10 The histogram of Ratio

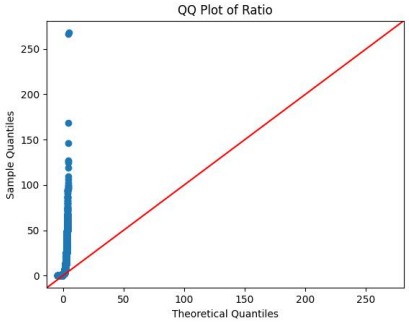

Figure 11 The Q-Q plot of Ratio

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

Table 6. Feature values of the three tests

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

Drawing the histogram shows the data distribution as a series of bars, which can intuitively show the concentration, dispersion degree and morphology of continuous variables. Theoretically, the histogram of normal distribution should be presented as a symmetrical bell-shaped curve, with the main features of symmetry, unimodal and tail behavior. Similarly, if the data follows a positive positive distribution, the linear line of the quantile of the sample data should be a 45° straight line on the Q-Q graph.

The histogram obtained by observation, the symmetry, unimodal and tail behavior characteristics of the three continuous variables are not obvious, and the Q-Q graph obtained by observation, the distribution of the three continuous variables is significantly deviated from the 45° straight line.

Table 6 shows the eigenvalues of the three continuous variables under the three tests, which obviously deviate significantly from the normal distribution.

Based on the above tests, Distance1, Distance2 and Ratio do not conform to the normal distribution

The obtained Pearson-correlation coefficient is as shown in Table 7:

	Distance1	Distance1	Distance2
Pearson correlation	0.1876	0.0919	0.4623

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

Although the Pearson correlation coefficient shows a certain linear relationship, because the results of the normality test indicate that these variables significantly do not conform to a normal distribution, the correlation analysis method should be used suitable to the non-normal distribution.

1) correlation test

Correlation tests for continuous variables

Because the continuous ϱvariable fields of the data set do not conform to the normal distribution and the linear relationship is poor, in order to test the correlation of the continuous variable fields and Fraud fields, it is necessary to choose sensitive to the non-linear relationship and do not require the normally distributed Spearman (Spearman) correlation coefficient test. Spearman's correlation coefficient is a non-parametric statistical metric used to assess the monotonic relationship between two variables. Spearman correlation coefficients convert the raw data into rankings and then calculate the correlation between ranks.

Spearman's correlation coefficiϱ $= 1 - \frac{6\sum d_i^2}{n(n^2-1)}$ent is calculated by the following formula: (

among,

Is the $d_i$ranking difference of each pair of observed values,

Isn the number of observed values.

The Spearman correlation coefficient is calculated by using the data of Articles 48 to 53 of Distance1 and Fraud:

Table 8. Raw data

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

Table 9 for the 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 10 The rank differences are calculatedd_i

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,

$$\sum d_i^2 = 9 + 1 + 9 + 1 + 16 = 36 \quad (2)$$

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

I. e., Spearman's correlation coefficient of the selected test set is-0.8.

* This result is only for the selected test set and is used to illustrate the working process of Spearman's correlation coefficient test and does not represent the complete set.

2) Correlation tests for categorical variables

Repeat, Card, and Pin Online Field belongs to categorical variables and shall adopt the test method suitable for categorical variables. The Chi-square test (Chi-squared Test) is a method used to analyze the association between two categorical variables, which determines the independence between variables by calculating the difference between the observed and expected frequencies. The chi-square test is a non-parametric test that does not require the data to follow a normal distribution or other specific distribution, which gives the chi-square test high flexibility in handling actual data, which is more reliable in dealing with larger sample sizes and more suitable for the case of this data set.

The chi-square test statistic $\chi^2$is calculated by the following formula:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

among,

- Is the $O_i$observed frequency (actual value)
- Is the $E_i$expected frequency (theoretical value)

The contingency tables were constructed based on the different values of the two categorical variables, which show the observed frequencies for each combination.

The expected frequency $E_i$is:

$$E_i = \frac{(R_i \times C_j)}{N} \quad (5)$$

Where, is the total number R_iC_j of frequencies in row i, is the total number of frequencies in

column j, and N is the total number of observations.

The contribution value of each cell was calculated $\chi^2\chi^2$ from the formula and summed to obtain the total statistics.

The calculation df formula of the degree of freedom is:

$$df = (r - 1) \times (c - 1) \quad (6)$$

This is the number rc of rows and the number of columns.

Based on the χ^2 statistics and degrees of freedom, find the chi-square distribution table or determine the p-value using statistical software. If the p-value is less than the significance level (usually 0.05), the null hypothesis is rejected that the field has no significant correlation with the Fraud field.

Model solution

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

Correlation analysis between continuous variables and telecom fraud

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

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

The Spearman correlation coefficient between Rratio and Fraud is 0.3428, indicating the moderate intensity positive correlation between the ratio of current transaction amount and last transaction amount and telecom fraud.

All of these correlation coefficients have a p-value of 0.0000, indicating a significant correlation.

Categorical variables and chi-square test results for telecom fraud

The chi-square value of Repeat and Fraud is 1.8278 and the p-value is 0.1764, indicating that whether trading in the same bank is not significantly associated with telecom fraud.

The card-square value of Card and Fraud is 3717.4490, and the p-value is 0.0000, indicating that whether the bank card is traded on the device is significantly related to telecom fraud.

The chi-square value of Pin and Flood is 10057.4125 and the p-value is 0.0000, indicating whether the PIN code is significantly associated with telecom fraud.

The chi-square value of Online and Flood is 36852.0237, and the p-value is 0.0000, indicating a significant association between online transactions and telecom fraud.

Conclusion

The correlation of the key variables is shown as follows:

The chi-square test of Repeat and Flood showed that trading in the same bank was not significantly associated with telecom fraud (p-value greater than 0.05).

Online and Fred's chi-square tests showed that online transactions are significantly associated with telecom fraud (p-value is much less than 0.05).

The correlation test shows that the five indicators of "transaction amount ratio", "whether to trade in the same bank", "whether to use the bank card to trade on the device", "whether to use PIN code" and "whether to trade online" have a strong correlation with whether telecom fraud occurs.

These results show that in the development measures to prevent telecom fraud, should focus on the transaction amount ratio (Ratio), whether to use the bank card on the equipment transactions (Card), whether to use PIN code (Pin) and whether online transactions (Online), because they have a strong correlation with telecom fraud, whether to trade in the same bank (Repeat), the impact of "telecom fraud occurs" is negligible.

Solution to question four

Data preprocessing

Problem 4 needs to establish a more complex prediction model. The complex model is extremely sensitive to the outliers of the data set, so it needs to pre-process the data set, select the outliers and replace them.

Outlier screening

Boxplot (Boxplot), a graphical representation method for showing the distribution of a set of data. It provides a visual overview of the data distribution by succinctly summarizing the median, quartile and outliers. The following are the main components of the boxplot and their representative statistical significance:

Components of the case chart:

Median (Median): median value of the data set, the bar of the box.

Quartiles (Quartiles):

Q1 (first quartile): the number at 25% of all values, the lower edge of the box.

Q3 (third quartile): Number 75% of all values, the upper edge of the box.

Quartile distance (Interquartile Range, IQR): the gap between Q3 and Q1, indicating the middle 50% of the data.

Shker (Whiskers): Line extending outside the box, usually to Q1-1.5 IQR and Q3 + 1.5 IQR, to define the normal data range. Points outside of this range are often treated as outliers or outliers.

Outliers (Outliers): These are those below Q1-1.5 IQR or above Q3 + 1.5 IQR.

The core point of the boxplot method to find outliers is to calculate the quartiles, considering values lower than Q1 n IQR or higher than Q3 + n IQR as outliers, where the standard value of n is 1.5.

The boxplot of the first seven fields of the Python code drawing is as follows:

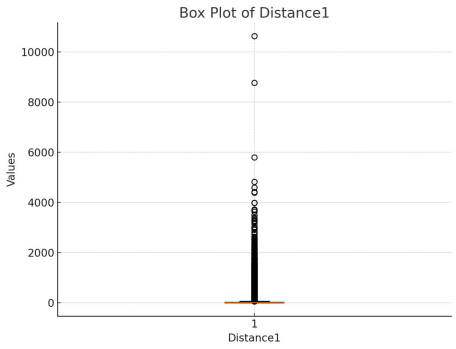

Figure 12 Distance1 boxplot

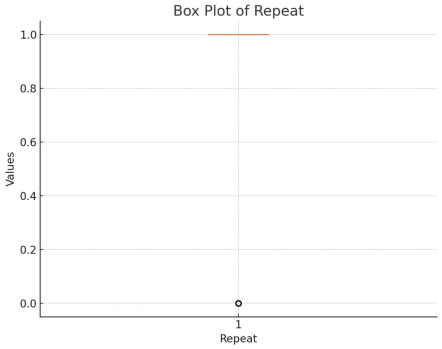

Figure 13 Repeat boxplot

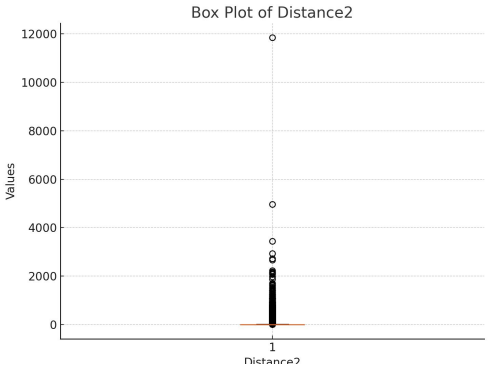

Figure 14 Distance boxplot

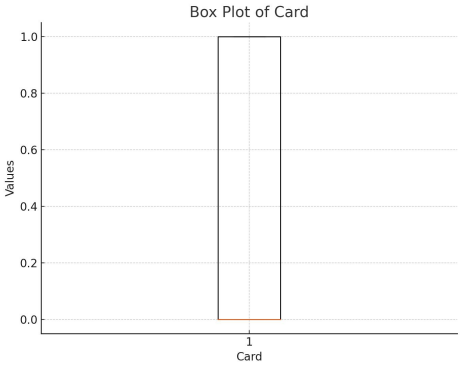

Figure 15 Card boxplot

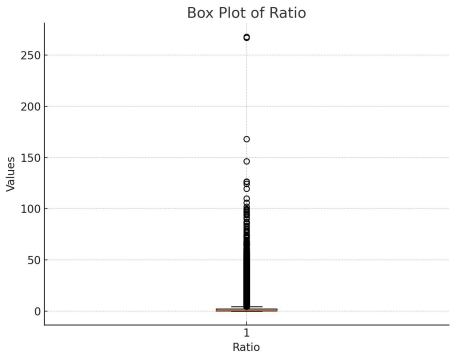

Figure 16 Box plot of Ratio

Taking n as the standard value of 1.5, the number of outliers of the first seven variables obtained by programming electrical calculation is shown in Table 11:

Table 11 Number of outliers for the top seven variables

Distance1	103631
Distance2	124367
Ratio	84386
Repeat	118464
Card	0
Pin	100608
Online	0

Fill with the k-nearest neighbor (k-NN) method

The k-nearest neighbor (k-Nearest Neighbors, k-NN) is a non-parametric method that performs classification, regression, or filling operations based on the similarity between neighboring points. When dealing with missing values or outliers, the k-NN fills up by finding the k nearest neighbors of a data point (i. e., the k points closest to the point in feature space), and then uses those neighbors to estimate or replace the missing or abnormal values.

flow of work:

Select k value: k is a hyperparameter representing the number of points closest to the target point in the data set.

Distance measure: Euclidean distance is usually used to determine the distance between data points, but Manhattan distance or other distance measures can also be used.

Find the nearest k neighbors: For each point in the dataset, find the nearest k data points in the feature space.

Calculation of fill values: For continuous variables, the mean or median of these k neighbors is usually used to populate the missing values; for categorical variables, use patterns (i. e. the most common category among these k neighbors).

In the practical application of this problem, the KNNImputer in Python is used to achieve the fast calculation, which automatically finds the k nearest neighbors of the points with outliers in each feature. If an eigenvalue is marked as abnormal by the boxplot method, the eigenvalue can be adjusted by considering the eigenvalue of its nearest neighbor.

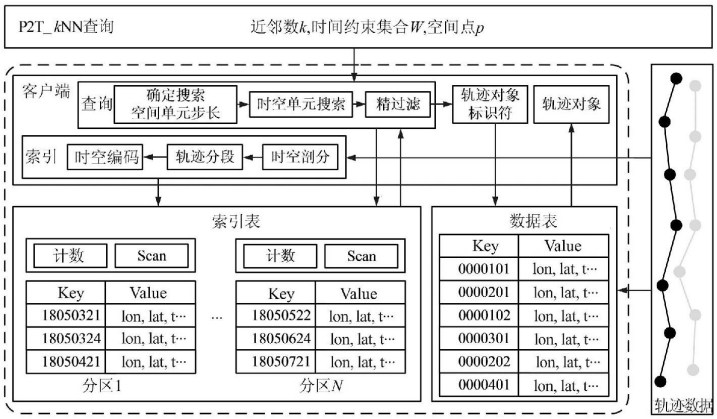

Figure 17 k-NN query processing framework

Predictive model of CNN-LTSM fusion based on attention mechanism

We created a prediction model of CNN-LTSM based on the attention mechanism, [8],

The model structure is provided as follows:

Input layer: receive transaction data, data shape depends on a specific time step and number of features.

CNN layer: The first layer is a convolutional layer, using the ReLU activation function, which is designed to capture the local characteristics in the transaction data.

The LSTM layer: receives the output from the CNN layer that helps the model understand the temporal dynamics in the data.

Attention layer: a custom Bahdanau attention layer, which enhances the focus of the model on important time steps by calculating the context vectors and attention weights.

Output layer: Use the full connection layer of the sigmoid activation function to predict whether the transaction is fraudulent.

Training and optimization

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

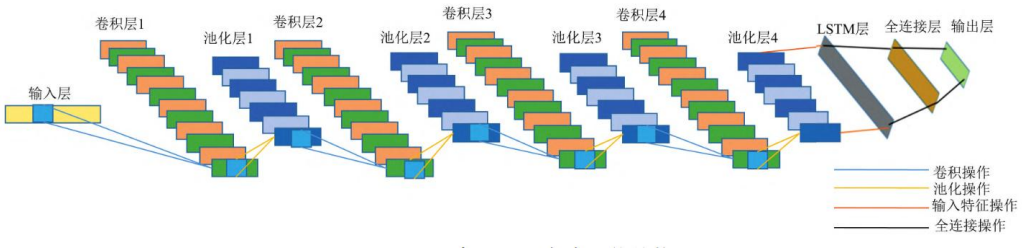

Figure 18: A CNN-LTSM composite grid structure

Convolutional neural network (CNN)

Process the input transaction data using a CNN (Convolutional Neural Network). CNN is a powerful neural network architecture, often used in image processing as well as any form of multi-dimensional array data processing, including time series and transaction data. The main purpose is to extract the local features in the data through the convolution operations.

The CNN layer mainly consists of the following components:

Convolution Layer: Convolution operations occur directly on the raw input data using a set of learnable filters (or convolution cores) that capture the local characteristics of the data. Each filter slides through the input data one by one, calculates the dot product of the filter and the data, and generates the output called the feature map (feature map).

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

* Pooling layer: In some CNN architectures, the pooling layer (such as maximum pooling) is used to further reduce the spatial dimension of the feature graph, strengthening the robustness of the model to small position changes.

In this prediction model, the use of the CNN layer can help to automatically identify complex patterns and features that may indicate fraud, without the manual design of feature extraction rules. The model first receives the transaction data as input, then automatically learns from the transaction data through its convolution kernel to help the features that to identify fraud, and finally outputs the feature graph to the next layer.

Long and Short Time Memory Network (LSTM)

In the field of deep learning, long-term short-term memory network (LSTM) is a special type of recurrent neural network (RNN), which is especially 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 disappearance problem that traditional RNN may encounter during training on long sequences.

Long and short time memory network (LSTM) is a special type of recurrent neural network (RNN), which is mainly used for tasks to process and predict sequence data. This model is used to capture the time series dependence in transaction data.

The core components of the LSTM include:

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

Forget gate (Forget Gate): deciding which information should be discarded from the cell state. Through a Sigmoid neural network layer control, which looks at the previous hidden state and the current input, it outputs a numerical value between 0 and 1, indicating how much old information is retained in each cell state.

Input gate (Input Gate): Update the cell state. First, a Sigmoid layer determines which values will be updated, and second, a Tanh layer creates a new vector of candidate values, which will be added to the state.

Cell state (Cell State): It is the "memory" part of the network, which runs through the whole chain, with only a slight linear interaction, so that information flows through the network without much change.

- Output gate (Output Gate): determines output based on the cell state. The output is the result

of the cell state being processed by tanh (squeezing values between-1 and 1) and then passing through a Sigmoid gate to determine which part of the information will be output.

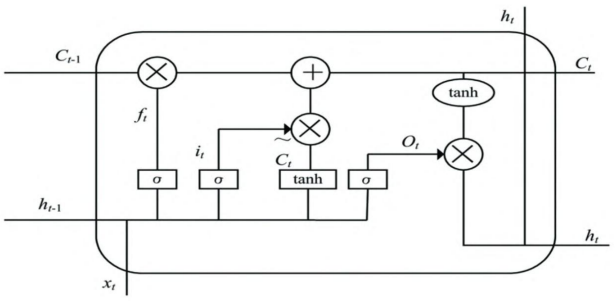

Figure 19 Long-and short-term cellular memory structure

In this prediction model, the LSTM layer undertakes the task of capturing the time-series dependence in the transaction data. LSTM receives the layer output features from the CNN layer, which represent the spatial properties of the transaction data at each time step. By maintaining the internal state (cell state), LSTM can remember and use the past information to help the model identify fraud patterns that may span multiple time steps. Its output for each time step provides a feature representation that integrates spatial characteristics and time dynamics.

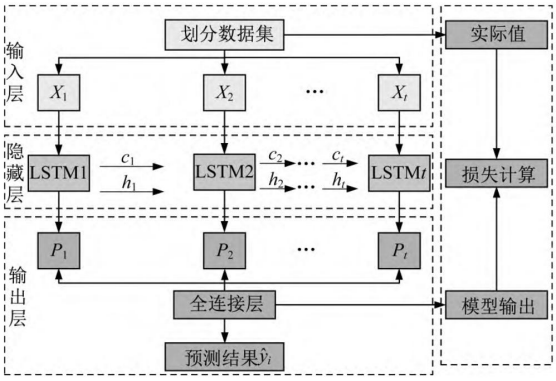

Figure 20 The LSTM grid structure

Attention attention mechanism

The built-in Attention layer of the Keras library implements a simple attention mechanism that is commonly used for sequence-to-sequence models, especially in natural language processing and time series analysis. This attention mechanism is implemented based on a weighted sum of attention as the input sequence, where the weights are determined by the relevant part of the sequence. The principle is as follows:

In the Attention layer of the Keras library, the core idea of the attention mechanism is to compute an attention score that indicates the importance of each part of the input sequence given the context of the current output. The calculation method typically involves the following characteristic values:

Query (Query): the representation of the current time step or the output of the previous layer.

Key (Key) and value (Value): Usually associated with the input sequence, keys are used to calculate the attention score, and values are actually weighted data.

Score calculation: calculate the relationship between query and key using dot product or other similarity measures.

Softmax: These scores are converted into probability values through the softmax layer to ensure that the sum of all the scores is 1.

Weighted sum: Use the above probability values (weights) to get the weighted sum, which is the output of the layer.

In this model, the application of the Attention layer can very directly enhance the ability of the model to focus on those time steps or characteristics that are most critical for identifying potential fraud. Specifically, each timestep of the model can be treated as a query, and the entire input sequence constitutes a set of keys and values.

The output of the LSTM serves as an input to the Attention layer, where the output of each time step can be treated as a query, while the output of the entire sequence is treated as both keys and values. The score is calculated by comparing the output (query) of each time step of the LSTM layer with the entire output sequence (keys). The scores were transformed using the softmax function, yielding a probability distribution representing the importance of each time step to the output of the current step. Eventually, these weights are used to create a weighted sequence output (the weighted sum of the values), which is enriched with information about the most important parts of the entire sequence.

The weighted output is then sent to the fully connected layer, using the activation function (Sigmoid function) to predict the output, namely whether the transaction is fraudulent.

Model Fusion —— Stacking (Stacking)

Stacking (Stacking) is an advanced model fusion technique that builds a new prediction model by combining the prediction results of multiple different models. This approach is based on an assumption that different models may learn different aspects of the data, and that by properly incorporating the outputs of these models, they can get better predictions than any single model.

Selection and training of the base model

CNN and LSTM were selected as base models for the following purposes:

CNN model: Focus on analyzing the spatial characteristics of transaction data, and capture abnormal patterns through the spatial attributes of transactions (such as amount size, transaction type, etc.).

LSTM model: Using its ability to process time-series data, to analyze the continuity and abnormal changes in trading behavior over time, for example, identifying abnormal increases in transaction frequency or unusual behavior within a specific time frame.

Each model was trained independently and optimized during training to extract the features that most favored the prediction.

2) Generate meta-features

After training the base models, they are used to predict the entire training set or the partial training set through cross-validation to generate meta-features.

CNN model output: fraud behavior feature diagram

The LSTM model output: a weighted sequence

3) Training of the meta-model

Use these meta-features as input to train a new model (called meta-models) to integrate the information of the base model and make the final predictions. The choice of meta-model can be logistic

regression, decision tree or any model suitable to handle the classification task. The steps for training the metamodel are as follows:

Input data: meta-feature, the output of the previous base model.

Target: the same target variable as in the base model training, that is, whether the transaction is a fraud.

Training: to train the meta-models to learn how to most effectively combine the predictions of the base model to optimize the accuracy of fraud detection.

The model is trained with a binary cross-entropy loss function using the Adam optimizer and may integrate early stop strategies to avoid overfitting, ensuring optimal performance on the validation data.

The division of the dataset

Data set division is a very critical step in machine learning, which can directly affect the training effect and generalization ability of the model. Its main purpose is to evaluate the ability of the model to generalize to new data, where the model can perform well on unseen data. In this study, the random method was used to divide the data set. Specifically, the train_test_split () function in the sklearn package was used to divide the size of the training set: test set =7:3. To ensure the repeatability of the results, random_state=42 was set.

Performance evaluation

The performance of the stacked model was evaluated, comparing the prediction of the model with the actual results, using indicators such as accuracy, recall, and F1 score.

Cross-validation: Cross-validation is used to ensure the robustness of the evaluation results and the generalization ability of the model.

Performance indicator calculation: Key performance indicators, such as accuracy, recall rate and F1 scores, are calculated to evaluate the performance of the model in different aspects.

Results analysis: analyze these indicators to determine the effect of the model in predicting the telecom bank card fraud, to determine whether there is overfitting or other problems.

The performance evaluation results are as follows:

Accuracy: 0.9966

Precision: 0.9773

Recall: 0.9842

F1 Score: 0.9808

The high accuracy (0.9966) indicates that the model also maintains good performance on the test set.

The high precision (0.9773) and high recall (0.9842) indicate that the model can accurately identify most of the fraud when predicting telecom fraud, while the false alarm rate is low.

The high F1 score (0.9808) considered precision and recall, indicating good overall performance of the model.

Model improvement versus the performance of the improved model

Introducing custom attention mechanisms

Bahdanau The attention mechanism improves the ability of neural network models to process

sequence data by paying different degrees of attention to different parts of the input. The core is creating a layer that can learn how to assign attention weights, which can be achieved in the following steps:

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

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

Softmax Layer: Apply the softmax function to the score and convert it into a probability distribution, which represents the importance of each input.

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

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

Add techniques to prevent overfitting

1） regularization

Dropout Is a common regularization technique that prevents neural network overfitting by randomly "discarding" (i. e., setting to zero) the activation values of some neurons during training. This approach forces the network to learn more robust features because it cannot rely on either neuron, because neurons may be removed randomly during training.

2) Batch normalization

Batch Normalization (Batch normalization) is another technique that accelerates the training process and improves the stability of the model by standardizing the input from the layer. It is achieved by reducing the internal covariate offset, applied to the activation function before each layer.

Performance evaluation of the improved model

The training procedure is set to train ten full iterations of the entire training set, during which the model will see the training data multiple times, each time with the opportunity to adjust its weight and bias to reduce prediction errors.

The process for each Epoch includes:

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

Calculate the loss: The difference between the model's predicted output and the actual value is calculated from the loss function. This loss represents how good the model has s current performance.

Backpropagation: calculate the gradient about each weight through the loss function, and then these gradients are used to update the weight of the model, which is completed by optimization algorithms (such as SGD, Adam, etc.).

Table 12 Model performance of the ten Epoch processes

	Training accuracy	Training loss	Validation accuracy	Validate the 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 proved that both training and verification accuracy are high, and the validation loss gradually decreases, indicating that the model gradually converges during training and has good generalization ability. From the fifth epoch, the validation accuracy is very close to 100%, and the validation loss is also very low, indicating that the model has been able to learn the patterns in the data very well.

Accuracy: 0.9984

Precision: 0.9933

Recall: 0.9888

F1 Score: 0.9911

The accuracy increased by 0.18 percentage points, and the high accuracy (0.9984) indicates that the model also maintained good performance on the test set.

The recall rate increased by 1.15 percentage points, and the high precision (0.9933) and high recall (0.9888) showed that the model can accurately identify most fraud when predicting telecom fraud, with a low false alarm rate.

The F1 score increased by 1.03 percentage points, and the high F1 score (0.9911) considered both precision and recall rates, indicating that the model performed well overall.

In addition, the confusion matrix is used to intuitively understand the prediction accuracy. The confusion matrix shows the performance of the model in each category, including true class (True Positives), false positive class (False Positives), true negative class (True Negatives), and false negative class (False Negatives). The confusion matrix provides intuitive visibility on which categories the model performs well and on which categories are problematic. Then draw the ROC curve, the ROC curve (receiver operation characteristic curve) and AUC (area under the curve) are important tools to evaluate the classification performance of the model, especially when the data set is unbalanced. The ROC curves are drawn by calculating the true rate (True Positive Rate, recall) and false positive rate (False Positive Rate) at different thresholds. The AUC value gives the ability of the model to distinguish positive classes from negative classes, and the higher the AUC value, the better the performance of the model. The AUC is calculated as the area under the ROC curve.

as illustrated in following figure:

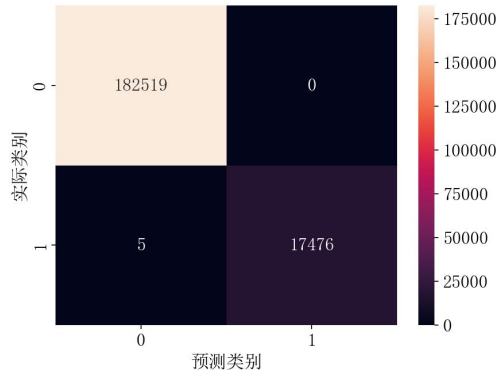

Figure 21 A confusion matrix plot of the model performance

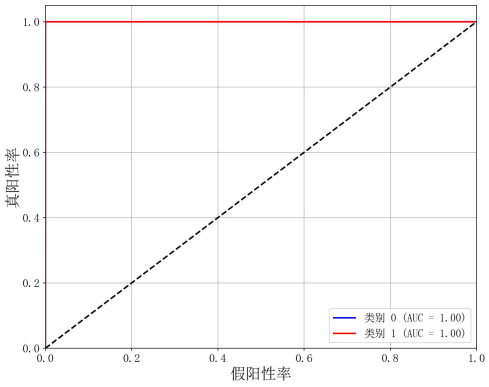

Figure 22 The ROC plot of the model performance

As can be seen from the confusion matrix, the model performs very well (false positive in avoiding misclassifying non-fraud categories as fraud (0), which means it is very accurate. However, five cases of fraud were wrongly judged as non-fraudulent, indicating that while the model's sensitivity (recall) is also high, there is still room for improvement. The AUC value is approximately 1, which indicates that the model has extremely strong classification power and can distinguish fraud from 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 still needs to verify its robustness on new or more diverse data. This performance of the model may be because the training data is very representative or somehow simplifies the real-world complexity, and perfect performance may sometimes mean overfitting, especially if the training data differ from the data in the actual application scenario.

Suggestions

Suggestions to the public security department

Focus on monitoring long distance transfer: Distance1 (the distance from the bank card transfer transaction site) and Distance2 (the distance from the last bank card transfer transaction) show a large positive SHAP value, indicating that the increase in distance may be related to telecom fraud. Public security departments should pay attention to abnormal long-distance transactions and cooperate with banks to monitor them.

Focus on abnormal transaction patterns: Ratio (the ratio of the amount of the card transfer transaction to the amount of the last card transfer transaction) has a large positive SHAP value, indicating that the significant change in the amount may be related to fraud. Analyof transaction patterns can identify unusual large transfers.

Strengthening data sharing: It is suggested that public security departments strengthen data sharing cooperation with banks and other financial institutions to ensure the timely transmission of fraud information. Use advanced data analysis techniques and artificial intelligence, such as the model shown, to analyze and predict potential fraud.

Improve the tracking efficiency of fraud cases: use the machine learning model to help the public security departments quickly identify the patterns and trends of fraud activities, and accelerate the tracking and crackdown of fraud gangs.

Advice to Banks

Strengthen online security measures: Online (whether it is an online bank card transfer transaction) SHAP value shows that online transaction is an important way of telecom fraud. Banks should strengthen network security, provide security tips for customers, and improve the security verification of online transactions.

Educating customers to identify fraud: For Card and Pin transactions, banks should better educate consumers about how to safely use bank cards and PIN codes, and how to identify possible fraud.

Strengthening transaction monitoring system: Banks should adopt advanced fraud detection systems, such as the model in this study, to monitor and analyze transaction behavior in real time, and quickly identify and block suspicious transactions.

Improve customer service response: Set up a quick response team to handle customer reports about suspicious transactions to ensure that customers get quick and effective help when they encounter problems.

Advice to citizens

Improve the alert to telecom fraud: educate citizens to identify the common means of telecom fraud, such as unsolicited online requests, abnormal account activity tips, etc.

Remote transfers are handled carefully: Citizens should also verify the authenticity of the transaction, especially in unusual cases.

Protect personal information: take good care of personal information, do not trust strange calls and text messages, do not click on unknown links at will.

Complex password and regular replacement: set the complex password and change it regularly, and do not share the bank card and password information with others.

Pay attention to safety tips: pay attention to the safety tips issued by banks and police, and keep abreast of the latest fraud methods and preventive measures.

Advantages and disadvantages of the model

Advantages of the model

High-dimensional data-processing capability. The use of logistic regression and ensemble learning methods allows the model to effectively process high-dimensional data and extract the key factors affecting the occurrence of telecom fraud. The stability and strong interpretability of logistic regression on feature selection are its significant advantages, while ensemble learning enhances the robustness of the results by integrating predictions from multiple models.

Dynamic analysis capabilities of the time-series data. Using the long-and short-term memory network (LSTM) to process time series data to capture the time dependence in transaction behavior, which is particularly important to understand the development pattern of fraud. LSTM is able to remember the long-term dependence information through its internal structure, which is more suitable for processing complex financial transaction data than the traditional time series analysis model.

Modeling power of characteristic nonlinear relationships. CNN is used to analyze and extract complex spatial features in transaction data, especially in showing its superiority in processing unstructured data. The CNN is able to automatically identify and utilize important features in images or sequence data that may be ignored when being manually extracted.

The ability to synthesize multiple data sources. The application of data fusion technologies, such as feature fusion and data integration, enables the model to comprehensively consider information from different sources and improve the accuracy of decision making. This is particularly critical for building a comprehensive telecom fraud detection system, as fraud usually involves multiple signals and data types.

Interpretative nature and transparency of the model. By using models such as logistic regression, it can provide interpretable model output, helping to understand which factors play a decisive role in fraud prediction. This is important for building trust and further optimizing the model.

Adaptability and scalability. The design of the model allows for the easy addition of new data sources and features, which is critical for changing strategies and approaches to combat telecom fraud. The scalability of the model ensures that the prediction performance can continue to improve as more data accumulates.

The shortcomings of the model

Computational resources consume costly. The CNN-LSTM model based on attention mechanism, which usually requires large computational resources for training and prediction, especially when processing large-scale datasets. This may limit the usefulness of the model in environments with limited computational power.

The model training time is long. The training process of composite models involves the training of multiple layer structures and multiple models, leading to a long overall training time. This may affect the iteration speed of the model, especially in application scenarios that require rapid response or frequent updates of the model.

Overfitting risk. Highly complex models such as deep learning models are prone to overfit on the training data, especially when the dataset is insufficient to cover all potential changes. Overfitting may lead to a less generalized model on new data.

The model is less explanatory. Although models such as logistic regression have good explanatory power, the internal decision-making mechanisms of deep learning models such as CNN and LSTM are relatively complex and difficult to interpret. This may be a disadvantage in cases where interpreting model predictions are needed for manual intervention or policy making.

Complexity of maintenance and updating. Your model contains multiple sub-models and complex structures, which make model maintenance and update work complex and time-consuming. In particular, in response to conceptual drift or data distribution changes, tuning and retraining the model may require a significant shift of effort.

High dependence on data quality and quantity. Model performance greatly depends on the quality and quantity of the training data. Insufficient or biased data may lead to insufficient model training, thus affecting the accuracy and reliability of the prediction.

Future expectations

Explore new model architectures and algorithms. Future studies could consider exploring emerging deep learning architectures, such as transformers (Transformers) and Graph Neural Networks (Graph Neural Networks), that have shown superior performance in handling complex data relationships and large-scale datasets.

Enhance the interpretation and transparency of the model. For highly sensitive applications such as telecom fraud detection, it is crucial to improve the interpretability of the models. Future research could

focus on the development and integration of new explanatory mechanisms, such as LIME (locally interpretable model-sensitive interpretation) or SHAP (SHapley Additive exPlanations), so that users and regulators can better understand the model's prediction and decision-making processes.

Realize real-time monitoring and prediction. With the development of technology, real-time data processing and prediction have become possible. Research can focus on developing dynamic models that can receive and analyze transaction data in real time and identify potential fraud in real time.

Cross-domain and cross-border data integration. Telecom fraud detection can be enhanced by integrating data from different areas (such as social media, geographic location data, etc.). Using these diversified data sources can help the model capture more complex fraud patterns and improve the accuracy and robustness of prediction.

Adaptive model to dynamic changes in data. Fraud strategies and patterns vary over time, so the development of models that can adapt to these changes is an important direction for future research. This may include leveraging online learning, incremental learning, or transfer learning techniques to continuously adapt and optimize models.

References

[1] Chen Xiaolei, Xu Hui. Construction of the evidence system for telecom fraud cases [J]. Network Security Technology and Applications, 2024 (5): 139-141.

[2] Qin Yutian. The logical development and crime control of the core crime of fraud [J / OL]. Journal of Taiyuan University (Social Science Edition), 2024,25 (4): 95-104.

[3] Zhang Hao. Analysis on ecological Management of "black and grey Industry Chain" of telecom network fraud crime [J]. Network Security Technology and Applications, 2024 (4): 148-150.

[4] Xiao Song, Huang Jianwu. First-order approximated knife modified ridge type estimation in a binary logistic regression model [J]. Journal of Hunan University of Arts and Sciences (Natural Science Edition), 2024,36 (2): 6-13.

[5] Zhang Jian, Liu Lin, Li Qian. M-Score model optimization study based on nonlinear logistic regression [J]. Business Accounting, 2024 (10): 83-86.

[6] Liao Wen. Research and Application of User Loss prediction Model of Internet Finance Based on Data Mining [D / OL]. South China University of Technology, 2021.

[7] Zhao Zeming. Research on intelligent habitat mode and behavior prediction based on data mining technology [D / OL]. Harbin Institute of Technology, 2021.

[8] Li Hao, Zhao Qing, Cui Chenzhou, et al. A stellar spectral classification algorithm [J] based on the composite depth model of CNN and LSTM. Spectroscopy and spectral analysis, 2024,44 (6): 1668-1675.

[9] Zhao Zeming. Research on intelligent 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 Communication Network Evaluation of Smart Substation [J]. Mechanical Design and Manufacture, 2024 (04): 28-32 + 37.

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

[12] Liu Shengjiu, Liang Shupeng, Liu Ying, et al. Analysis and application 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 the XGBoost and LR fusion model [J]. Journal of Chongqing University of Technology (Natural Science), 2024,38 (03): 195-200.