ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ
МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ
INTERNATIONAL UNIVERSITY

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2023 (15) 3

Шілде – қыркүйек

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншіктенуші: «Халықаралық ақпараттық технологиялар университеті» АҚ (Алматы қ.).
Қазақстан Республикасы Ақпарат және әлеуметтік даму министрлігінің Ақпарат комитетінде – **20.02.2020** жылы берілген.
№ KZ82VPY00020475 мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: ақпараттық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологияларға арналған.
Мерзімділігі: жылына 4 рет.
Тиражы: 100 дана
Редакцияның мекенжайы: 050040, Алматы қ-сы, Манас к-сі, 34/1, 709-кабинет, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

---

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

МАЗМҰНЫ

ӘЛЕУМЕТТІК-ЭКОНОМИКАЛЫҚ ЖҮЙЕЛЕРДІ ДАМЫТУДАҒЫ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР

АҚПАРАТТЫҚ ТЕХНОЛОГИЯЛАР

СОДЕРЖАНИЕ

ЦИФРОВЫЕ ТЕХНОЛОГИИ В РАЗВИТИИ СОЦИО-ЭКОНОМИЧЕСКИХ СИСТЕМ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

CONTENTS

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 4. Is. 3. Number 15 (2023). Pp. 104–118
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2023.15.3.010

UDC 004.62

PREDICTING FRAUDULENT CRYPTOCURRENCY TRANSACTIONS USING LOGISTIC REGRESSION

I.L. Khlevna*, V.O. Deineha

Iuliia L. Khlevna — Doctor of Technical Sciences, Associate Professor of the Department of Technology Management
ORCID: 0000-0002-1807-8450. E-mail: yuliia.khlevna@knu.ua;
Vladyslav O. Deineha — Master Student of the Department of Technology Management, Taras Shevchenko National University of Kyiv
ORCID: 0009-0008-3123-2302. E-mail: deinehav@fit.knu.ua.

Abstract. The article conducts a SWOT analysis of cyber threats, identifies the strengths and weaknesses of cryptocurrencies. It is determined that the development of cryptocurrencies forms a new class of digital assets, which is attracting increasing attention from the economic and financial communities and information technology. The issue of detecting fraudulent transactions with cryptocurrency is highlighted. To solve the problems of detecting fraudulent transactions, the authors propose to use new technologies based on data analysis methods, in particular, the development of logistic regression models. The following algorithm for classifying fraudulent transactions with cryptocurrency is proposed, which is reduced to the classical data classification scheme. The following steps are highlighted: data loading into the dataset and primary analysis, data preparation for analysis, division into training and test samples, application of the classification algorithm on the training sample, evaluation of the model accuracy on the test sample, model optimization if necessary, and conclusion, where, if the model accuracy is high, it can be used to classify fraudulent transactions. The main task of the presented stages is to detect suspicious cryptocurrency transactions with as few false positives as possible. The classification of cryptocurrency transactions is proposed to be carried out with the Ethereum cryptocurrency. The R language and its integrated processing environment R Studio are chosen as tools. A logistic regression model has been developed to detect fraudulent transactions with cryptoassets. The model checks a new transaction for fraud. The model's high accuracy of 98 percent demonstrates its effectiveness. The model can be improved to take into account new types of fraudulent

schemes and applied to analyze transactions with different assets, making it promising for use in financial institutions and cryptocurrency exchanges.

Keywords: crypto market, logistic regression, predicting, classification, data analysis

For citation: I.L. Khlevna, V.O. Deineha. Predicting fraudulent cryptocurrency transactions using logistic regression//INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES. 2023. Vol.4. No.3. Pp.104–118 (In Eng.). https://doi.org/10.54309/IJICT.2023.15.3.010

ЛОГИСТИКАЛЫҚ РЕГРЕССИЯНЫ ҚОЛДАНА ОТЫРЫП, АЛАЯҚТЫҚ КРИПТОВАЛЮТА ОПЕРАЦИЯЛАРЫН БОЛЖАУ

И.Л. Хлевна*, В.О. Дейнега

Хлевна Юлия Леонидовна — техника ғылымдарының докторы, доцент, Тарас Шевченко атындағы Киев ұлттық университетінің басқару технологиялары кафедрасының профессоры
ORCID: 0000-0002-1807-8450. E-mail: yuliia.khlevna@knu.ua;
Дейнега Владислав Александрович — Тарас Шевченко атындағы Киев ұлттық университетінің басқару технологиялары кафедрасының магистрі
ORCID: 0009-0008-3123-2302. E-mail: deinehav@fit.knu.ua.

Аннотация. Мақалада киберқауіптердің SWOT талдауы жүргізіледі, криптовалюталардың күшті және әлсіз жақтары анықталады және алаяқтық криптовалюта транзакцияларын анықтау мәселесі қарастырылады. Мәселелерді шешу үшін авторлар деректерді талдау әдістеріне негізделген жаңа технологияларды, атап айтқанда логистикалық регрессия модельдерін әзірлеуді ұсынады. Алаяқтық криптовалюта транзакцияларын жіктеу алгоритмі ұсынылады, ол классикалық деректерді жіктеу схемасына дейін азаяды. Келесі қадамдар ерекшеленеді: деректерді деректер жиынтығына жүктеу және бастапқы талдау, деректерді талдауға дайындау, оқыту және тест үлгілеріне бөлу, оқыту үлгісіне жіктеу алгоритмін қолдану, сынақ үлгісіндегі модельдің дәлдігін бағалау, қажет болған жағдайда модельді оңтайландыру және егер модель болса, қорытынды дәлдік жоғары, оны алаяқтық операцияларды жіктеу үшін пайдалануға болады. Ұсынылған кезеңдердің негізгі міндеті-күдікті криптовалюта операцияларын мүмкіндігінше аз жалған позитивтермен анықтау. Cryptocurrency транзакцияларының жіктелуі Ethereum cryptocurrency көмегімен ұсынылады. Құралдар ретінде R тілі және оның интеграцияланған R studio өңдеу ортасы таңдалады. Алаяқтық криптоактивті транзакцияларды анықтау үшін логистикалық регрессия моделі жасалды. Модель жаңа транзакцияны алаяқтық үшін тексереді. 98 пайызды құрайтын модельдің жоғары дәлдігі оның тиімділігін көрсетеді. Модель алаяқтық схемалардың жаңа түрлерін ескере отырып жетілдірілуі мүмкін

және әртүрлі активтермен транзакцияларды талдау үшін қолданылуы мүмкін, бұл оны қаржы институттары мен криптовалюта биржаларында пайдалану үшін перспективалы етеді.

Түйін сөздер: крипто нарығы, логистикалық регрессия, болжау, жіктеу, деректерді талдау

Дәйексөз үшін: И.Л. Хлевна, В.О. Дейнега. Логистикалық регрессияны қолдана отырып, алаяқтық криптовалюта операцияларын болжау//Ақпараттық және коммуникациялық технологиялардың халықаралық журналы. 2023. V.4. № 3. Бет 104–118 (ағылшын тілінде). https://doi.org/10.54309/IJICT.2023.15.3.010

ПРОГНОЗИРОВАНИЕ МОШЕННИЧЕСКИХ ТРАНЗАКЦИЙ С КРИПТОВАЛЮТОЙ С ИСПОЛЬЗОВАНИЕМ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

И.Л. Хлевна, В.О. Дейнега

Хлевна Юлия Леонидовна — доктор технических наук, доцент, профессор кафедры технологий управления Киевского национального университета имени Тараса Шевченка

ORCID: 0000-0002-1807-8450. E-mail: yuliia.khlevna@knu.ua;

Дейнега Владислав Александрович — магистр кафедры технологий управления Киевского национального университета имени Тараса Шевченка

ORCID: 0009-0008-3123-2302. E-mail: deinehav@fit.knu.ua.

Аннотация. В статье проводится SWOT-анализ киберугроз, выявляются сильные и слабые стороны криптовалют и освещается проблема обнаружения мошеннических транзакций с криптовалютой. Для решения задач авторы предлагают использовать новые технологии, основанные на методах анализа данных, в частности, разработку моделей логистической регрессии. Предлагается алгоритм классификации мошеннических транзакций с криптовалютой, который сводится к классической схеме классификации данных. Выделяются следующие этапы: загрузка данных в набор данных и первичный анализ, подготовка данных для анализа, разделение на обучающую и тестовую выборки, применение алгоритма классификации к обучающей выборке, оценка точности модели на тестовой выборке, оптимизация модели при необходимости и заключение, где, если модель точность высока, ее можно использовать для классификации мошеннических транзакций. Основная задача представленных этапов – обнаружить подозрительные транзакции с криптовалютой с как можно меньшим количеством ложных срабатываний. Классификацию криптовалютных транзакций предлагается осуществлять с помощью криптовалюты Ethereum. В качестве инструментов выбран язык R и его интегрированная среда обработки R Studio. Для выявления мошеннических транзакций с криптоактивами была разработана

модель логистической регрессии. Модель проверяет новую транзакцию на предмет мошенничества. Высокая точность модели, составляющая 98 процентов, свидетельствует о ее эффективности. Модель может быть усовершенствована с учетом новых типов мошеннических схем и применена для анализа транзакций с различными активами, что делает ее перспективной для использования в финансовых учреждениях и на криптовалютных биржах.

Ключевые слова: крипторынок, логистическая регрессия, прогнозирование, классификация, анализ данных

Для цитирования: И.Л. Хлевна, В.О. Дейнега. Прогнозирование мошеннических транзакций с криптовалютой с использованием логистической регрессии//Международный журнал информационных и коммуникационных технологий. 2023. Т. 04. № 3. Стр. 104–118 (На анг.). https://doi.org/10.54309/IJICT.2023.15.3.010

Introduction

The emergence of digital money is closely linked to the development of technologies, including cryptography and blockchain. Cryptocurrencies have become a new class of digital assets that is attracting increasing attention from the economic and financial communities and information technology. They share common features with traditional assets, but also have their own concept.

One of the main features of cryptocurrency is that it acts as a virtual currency. The owner of such a cryptocurrency stores it on his computer or in a smartphone application in a so-called "wallet" that only he has access to. The concept of cryptocurrencies is also that there is no regulator in circulation. Thus, there is no cryptocurrency central bank that could decide, for example, to increase the supply of cryptocurrency and thus reduce its value. The value is in the hands of the free market. Cryptocurrency trading takes place electronically, without the involvement of any banking system, directly between cryptocurrency users. This means that the transaction is not controlled in any way. Fraudulent transactions are becoming increasingly common in the crypto asset community. Cybercriminals use various methods, such as phishing attacks, wallet hacking, and other fraudulent schemes, to gain unauthorized access to cryptocurrencies. In addition, cryptocurrencies are also at risk of hacking and theft, which leads to the loss of digital assets in significant amounts. Such problems lead to a loss of confidence in cryptocurrencies as a means of investment and payment. This can lead to a decrease in investor interest and a decrease in the volume of cryptocurrency trading. Most of the security methods currently in use are aimed at ensuring the safety of cryptocurrency storage, such as the use of cold wallets and two-factor authentication. However, in order to ensure the security of cryptocurrency transactions, new technologies need to be developed, based on data processing and information technology. Technologies will help prevent fraud and protect users from theft of digital assets.

One solution may be to use blockchain technology, which is the basis for most cryptocurrencies. Blockchain technology is a distributed database that stores information about transactions in the form of a chain of blocks. Each block contains information

about the previous block, which makes it impossible to make changes to an existing transaction. This helps to ensure the security and transparency of transactions, as well as protect users from fraudulent schemes.

In addition, the development and use of new technologies, such as machine learning and artificial intelligence, can also help combat cryptocurrency fraud. For example, data analytics and anomaly detection can help identify fraudulent schemes and prevent unauthorized transactions.

In this regard, the article identifies the scientific and applied task of analyzing and forecasting fraudulent transactions with cryptocurrency. Development of new methods for protecting cryptocurrency transactions, based on data analysis, artificial intelligence and information technology, is a prerequisite for ensuring confidence in cryptocurrencies as a means of investment and payment, as well as for further growth and development of the cryptocurrency market.

Materials and methods

Literature analysis of fraudulent operations with cryptocurrency

Every year, attention to cryptocurrencies is growing and more and more scientific papers are appearing on the development and prospects of the crypto market. In turn, this is due to the development of technologies, in particular, data analysis methods, artificial intelligence and information technology.

In this article (Sabry, 2020), the researchers consider the application of artificial intelligence methods to solve problems related to cryptocurrencies, such as price forecasting, trends, volatility, portfolio construction, and fraud detection. The article is an overview, which helps to navigate the huge number of different studies that apply AI methods in the field of cryptocurrencies and to highlight open areas that require future development in this very dynamic field. The review can help researchers interested in applying AI and machine learning techniques to the field of cryptocurrencies by providing simplified overviews of some of the research conducted in this area and the methods used, as well as listing some of the available datasets that they have used to solve various cryptocurrency problems. The paper is broad in scope and does not cover the main machine learning methods in the fraud domain.

The study of cryptocurrencies as a subject of fraud is reflected in the paper (Kutera, 2022). The study showed that cryptocurrencies are a new field of research, although they are developing quite intensively. The main areas of research include types of cryptocurrency fraud, methods of crime detection, risks associated with blockchain technology, money laundering, and legal regulation of cryptocurrencies. Money laundering is currently the most common type of fraud. It was also noted that pyramid schemes are the second most common fraud. It is possible to study not only existing but also undisclosed research trends. The author identifies the key types of fraud that lead to the most significant financial losses and identifies areas for further research that have profound practical implications for market participants. This paper provides a basis for forming a fraud detection model based on the proposed classification.

This is partially reflected in a publication (Bartoletti, 2021) that describes a study of cryptocurrency fraud, which includes the creation of a collection of fraudsters, the

development of a fraud classification tool, and the analysis of the distribution and correlation between fraud types. The authors note that many fraudulent activities do not rely on the functions of a particular blockchain, but simply use the blockchain's own cryptocurrency as a means of payment, and provide recommendations for countering crypto fraud based on their experience. However, the paper does not present the sequence of data processing.

The paper (Ismail Alarab, 2022) continues with a report that, as a result of data preprocessing, the models used on the Bitcoin and Ethereum datasets perform better than their original models. In addition, it was found that the most effective learning algorithms perform even better after reducing the number of features in these datasets compared to their initial studies.

The application of logistic regression as a machine learning method used to solve classification problems where it is necessary to determine which class an object belongs to is reflected in the paper (Jiashi Feng, 2014). It is widely used in various industries where it is necessary to make decisions based on certain features. This paper (Mohammed, 2022) proposes a method for avoiding credit card fraud using logistic regression. Therefore, it is of interest to study fraudulent transactions in the cryptocurrency market using logistic regression.

In addition, the literature presents other methods that can be used to classify fraudulent transactions. In the paper (Aponte-Novoa, 2022), the authors explore several machine learning models for classifying cryptojacking on websites, including logistic regression, decision tree, random forest, gradient boosting classifier, k-nearest neighbors, and XGBoost. The authors conclude that simple models such as logistic regression, decision tree, random forest, gradient boosting classifier, and k-nearest neighbors models can achieve a level of success similar to or even superior to advanced algorithms such as XGBoost and even other deep learning-based works. In addition, the authors note that the simplicity of these models helps the evaluator interpret the results and understand the inner workings of these models compared to advanced models that are considered black boxes. The study (Wang, 2022) found that in their case, LR demonstrates greater efficiency in predicting credit card fraud compared to the support vector machine K-nearest neighbor (KNN) and decision tree (DT) classifiers.

Thus, analyzing fraudulent cryptocurrency transactions is a necessary tool to ensure transaction security and prevent fraud. Various methods and tools for analysis allow for the detection of fraudulent transactions, which in turn helps to fight money laundering, terrorist financing, and other criminal activities related to cryptocurrencies.

However, despite the fact that fraud analysis is an effective tool, it cannot guarantee 100% protection against fraud. Fraudsters are constantly implementing new tricks and coming up with new schemes that can bypass the security system.

As the analysis has shown, despite the scientific and practical results obtained in the field of analyzing fraudulent transactions with cryptocurrencies, the issue requires the development and development of technology for detecting fraudulent transactions based on data analysis methods, in particular, the development of logistic regression models.

Research methods include mathematical apparatus, models and methods of statistical analysis, including regression analysis. The R language and its integrated processing environment R Studio were chosen as tools.

Results and discussion

Building a technology for predicting fraudulent transactions in cryptocurrencies

Using the SWOT analysis of cyber threats, we identified the strengths and weaknesses of cryptocurrencies. Among the strengths, we can identify the fact that cryptocurrencies are decentralized, for example, bitcoin is not controlled by the state or any private company. Open validity, which means that you can determine the authenticity of a transaction at any time. A wide range of different cryptocurrencies. An uninterrupted and fault-tolerant network, which is the very essence of crypto and stems from its decentralization. Independence from the geographical location of the entities transferring the currency. A high level of protection against hacking is associated with the cryptographic nature of cryptocurrencies.

Weaknesses include the lack of control over money transfers, as the network is decentralized, there is no supervisory authority that can monitor and control suspicious transfers. High volatility, the cryptocurrency rate is unpredictable and influenced by many factors. High risks in the form of password loss and the possibility of theft, as there is no regulatory body. Limited availability, as not everywhere can accept cryptocurrencies as a means of payment, and not everywhere cryptocurrencies are allowed.

Among the options, we can highlight such as distribution as a means of alternative to traditional payments, due to decentralization, and the lack of a regulatory body. Due to the anonymity of the increase in possible payments, conditional charitable donations if the sender wants to remain anonymous. Conducting transactions regardless of the boundaries. A reliable way to store cryptocurrencies passively, provided they are properly protected. Openness and transparency, as a likely option, simplifies audit activities for organizations.

Possible threats include difficult forecasting for medium-term currency storage. A sharp change in trends, the possibility of a sharp drop in the crypto rate. Possible emergence of pyramid schemes and certain similarities with them. Possibility of attacks by holding a controlling stake in all generating capacities, monopolization. Some risks with certain cryptocurrency storage.

The algorithm for solving the problem of classifying fraudulent transactions is reduced to the algorithm for building a classification model, which is proposed to be implemented according to the classical scheme shown in Figure 1.

Step 1: Load the data into the dataset and conduct an initial data analysis. In the study, the data was taken from a CSV file from the kaggle website.

Step 2: Prepare the data for analysis. This may include handling missing values, normalization, outlier removal, etc.

Step 3: Split the data into training and test samples. The ratio in the practical part was chosen to be 60 percent training to 40 percent test.

Steps 4-5: Apply the classification algorithm on the training set using the chosen algorithm, for example, logistic regression was used in this work.

Step 6: Evaluate the accuracy of the model on the test set. This can be done using metrics such as accuracy, completeness, F-measure, etc.

Step 7: Optimize the model, if necessary, and repeat steps 4–6.

Step 8: Conclusion. If the model works with high accuracy, it can be used to classify fraudulent cryptocurrency transactions. If the accuracy is low, you need to optimize the model or use a different classification algorithm.

Building a predictive model to detect illegal cryptocurrency transactions.

It is proposed to classify cryptocurrency transactions with the Ethereum cryptocurrency. For the study, the main data will be taken from the kaggle website (Ethereum Fraud Detection Dataset), the dataset is unbalanced.

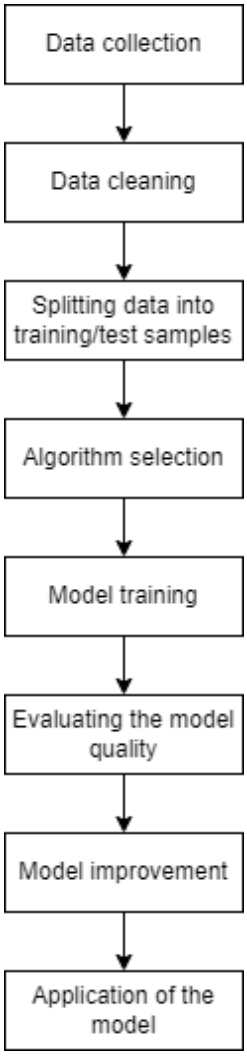

Figure 1 - Scheme of the algorithm for building a classification model

Let's go through the algorithm step by step regarding the practical part of the study.

Based on the analysis of the dataset, the following data metrics are identified:

Index: line index number

Address: the address of the Ethereum account

FLAG: whether the transaction is fraudulent or not

Avg min between sent tnx: average time between sent transactions for the account in minutes

Avgminbetweenreceivedtnx: average time between received transactions for the account in minutes

TimeDiffbetweenfirstand_last(Mins): time difference between the first and the last transaction

Sent_tnx: the total number of sent normal transactions

Received_tnx: the total number of received regular transactions

NumberofCreated_Contracts: total number of created contract transactions

UniqueReceivedFrom_Addresses: the total number of unique addresses from which the account received transactions

UniqueSentTo_Addresses20: the total number of unique addresses from which the account sent transactions

MinValueReceived: the minimum value on the ether that has ever been received

MaxValueReceived: the maximum value in the ether that has ever been received

AvgValueReceived5 Average value ever received on the ether

The logistic regression method was chosen as a classification method, which is well suited to our data. The R language and its integrated processing environment R Studio were chosen as tools. We will build a model that describes the relationship between the input variables and the probability of belonging to a certain class using logistic regression. Usually, this probability is expressed using the logistic distribution function (sigmoid function), which has the formula (Logistic Regression):

$p(x) = 1 / (1 + \exp(-x))$

where p(x) is the probability of an object belonging to a class, x is a linear combination of input features and their weights. The weights and the intersept (free term) are determined in the process of training the model on the training dataset.

To evaluate the accuracy of the model, it is proposed to use the loss function (LogLoss), which measures the discrepancy between the actual and predicted values. The task is to minimize the loss function using the gradient descent method or other optimization methods.

It is determined that to build a fraud detection model in R, the caret (Classification And REgression Training) package in the R programming language is needed, which provides a wide range of tools and functions for training and evaluating machine learning models. The caret package provides a wide range of machine learning algorithms for building models, including linear regression and many others.

This package will be valuable for fraud detection in terms of simplifying the process of building and evaluating models, selecting features, tuning hyperparameters, and evaluating model quality.

Import data:

```
d <- read.csv("transaction_dataset.csv", header = T, sep = "," )
Removed unnecessary columns:
x <- d[,-c(1,2,3,ncol(d)-1,ncol(d) )]
```

The data obtained are shown in Figure 2.

	FLAG	Avg.min.between.sent.tnx	Avg.min.between.received.tnx	Time.Diff.between.first.and.last..Mins.	Sent.tnx	Received.Tnx	Number.of.Crea
1	0	844.26	1093.71	704785.63	721	89	
2	0	12709.07	2958.44	1218216.73	94	8	
3	0	246194.54	2434.02	516729.30	2	10	
4	0	10219.60	15785.09	397555.90	25	9	
5	0	36.61	10707.77	382472.42	4598	20	
6	0	9900.12	375.48	20926.68	2	3	
7	0	69.46	629.44	8660.35	25	11	
8	0	1497.39	176.84	319828.05	213	5	
9	0	0.00	0.00	496.62	1	1	
10	0	2570.59	3336.01	30572.70	8	3	
11	0	32.45	12921.57	129540.15	10	10	
12	0	3716.41	1448.09	385961.98	8	246	
13	0	0.00	12431.27	198900.25	0	16	
14	0	9520.70	5776.32	78197.58	7	2	
15	0	14106.66	3742.82	540061.90	32	24	
16	0	757.91	11.08	25802.32	34	3	
17	0	3.13	4923.24	280803.43	57	57	

Showing 1 to 17 of 9,841 entries, 46 total columns

Figure 2 - Table of pre-cleaned data

A closer look at the table's columns reveals that most of the data has a lognormal distribution.

Figures 3 and 4 show the histogram of the second column of the table. You can see the lognormal distribution.

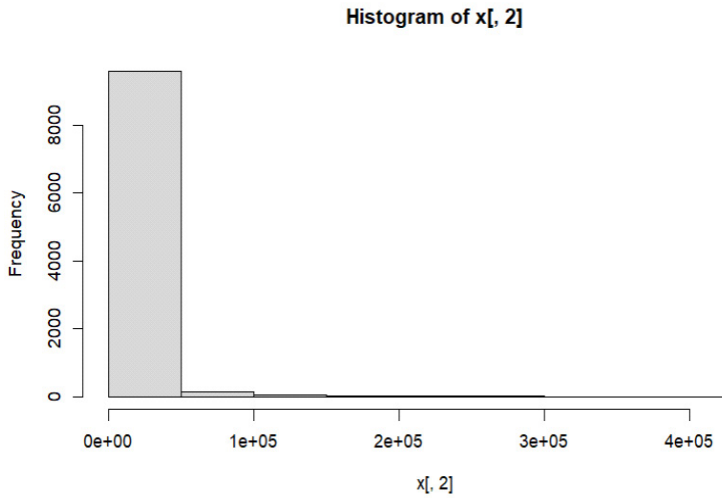

Figure 3 - Histogram in lognormal distribution

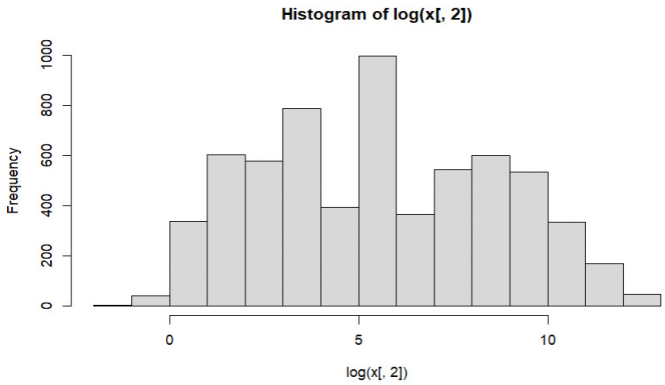

Figure 4 - Histogram after the logarithmization process

Perform the processing by logarithmizing the required columns while preserving zero during the process:

```
r <- 1:46
r<- r[-c(1,7,23,27,32,33,34,35,42,43,44)]
for (i in r) {
  x[,i] <- log(x[,i])
  x[!is.finite(x[,i]),i] <- 0
}
```

Some outliers and empty columns have been removed:

```
x <- x[x[,7]<=1,]
x[,c(16,17,18,22,27,23,30,32,33,34,35,42,43,44)] <- NULL
```

The dataset has been rebalanced. This is because the dataset is unbalanced:

To do this, the general table is divided into two, according to the flag variable, which is an indicator of the validity of the transaction.

```
x0 <- x[x$FLAG==0,]
x1 <- x[x$FLAG==1,]
xe <- x1
xe[] <- NA
```

Figures 5 and 6 show the number of rows in the table with the corresponding flag.

Showing 1 to 17 of 7,631 entries, 32 total columns

Figure 5 - Number of rows in a table with a zero flag

Showing 1 to 17 of 2,174 entries, 32 total columns

Figure 6 - Number of rows in a table with flag one

The next step was to fill the rows of the table where the total number of rows is inferior to the comparable capacity potential. We will use the method of filling the row with average values.

```
m <- colMeans(x1)
for (i in 1:ncol(x1)) {
  xe[,i] <- m[[i]]
}
xe <- as.data.frame(lapply(xe, rep, 3))
x1 <- xe
```

Based on the previous data processing, the transaction category ratio is more balanced. As a result, it is worth combining the two tables into one common table:

```
x <- rbind(x0,x1)
```

Removing highly correlated predictors:

To remove strongly correlated predictors, we will use the following method:

```
for (i in 2:ncol(x)) {
  for (j in 2:ncol(x)) {
    if(cor(x[,i],x[,j])>0.80 && i!=j){
      cat(" Correlation between ",i," and ",j," is ",cor(x[,i],x[,j]),"; \n")
    }
  }
}
x[,c(3,6,11,14,15,17,19,20,21,22,23,24,26,29)] <- NULL
```

After analyzing the data obtained after preprocessing, we can move on to the next step - modeling.

Modeling:

The flag column was set as a factor variable.

```
x$FLAG <- as.factor(x$FLAG)
```

To perform the modeling, the data frame was divided into two parts: a part of data for model training and a part of test data for model validation, respectively 60 percent to 40 percent. A function from the caret package was used for this purpose.

```
xsplit <- createDataPartition(x$FLAG, p = 0.6, list = FALSE)
trainx <- x[xsplit,]
testx <- x[-xsplit,]
```

To calculate the logistic regression, we used the train function, where we specified method = glm in the parameter, which indicates a general linear model. It is also necessary to specify the family = binomial option, which indicates that we want to use logistic regression.

```
model <- train(FLAG~., data = trainx, method="glm", family = "binomial")
```

After training, evaluate the characteristics of the resulting model (Fig. 7).

```
model
```

```
Generalized Linear Model

8493 samples
  17 predictor
   2 classes: '0', '1'
```

```
No pre-processing
Resampling: Bootstrapped (25 reps)
Summary of sample sizes: 8493, 8493, 8493, 8493, 8493, 8493, ...
Resampling results:

  Accuracy   Kappa
  0.9888712  0.977655
```

Figure 7 - Characteristics of the resulting model after training

The previously obtained test data was used to evaluate the effectiveness of the logistic regression model

pr <- predict(model, newdata = testx)

The accuracy assessment is based on the confusion matrix. A confusion matrix is a table in which predictions are categorized against actual values. It includes two dimensions, with one representing the predicted values and the other the actual values (Fig. 8).

cmat <- confusionMatrix(pr,testx$FLAG)

cmat

```
          Confusion Matrix and Statistics

                    Reference
          Prediction    0    1
                   0 2966    0
                   1   86 2608

                        Accuracy : 0.9848
                          95% CI : (0.9813, 0.9878)
             No Information Rate : 0.5392
             P-Value [Acc > NIR] : < 2.2e-16

                           Kappa : 0.9695

          Mcnemar's Test P-Value : < 2.2e-16

                      Sensitivity : 0.9718
                      Specificity : 1.0000
                   Pos Pred Value : 1.0000
                   Neg Pred Value : 0.9681
                       Prevalence : 0.5392
                   Detection Rate : 0.5240
             Detection Prevalence : 0.5240
                Balanced Accuracy : 0.9859

                 'Positive' Class : 0
```

Figure 8 - Confusion matrix and model statistics

As a result, a logistic model was built that produces an accuracy of approximately 98 percent on test data.

A model based on logistic regression is widely used for data analysis and classification tasks. It allows you to predict the probability of an object belonging to a certain class and, therefore, can be used to detect fraudulent transactions. The model's high accuracy of 98 percent indicates that it is highly effective, similar to other studies (Mohammed, 2022; Aponte-Novoa, 2022) that used logistic regression on their data. This means that the model can be used in real-world settings to detect fraudulent transactions with high accuracy, which is an important result in the fight against cyber threats and financial crime.

A model for predicting fraudulent transactions in cryptocurrencies can be of great importance to financial institutions and cryptocurrency exchanges, which can use it to prevent financial losses and improve transaction security. Prospects for further research are to improve the model to take into account new types of fraudulent schemes and increase its accuracy. Also, the model can be used to analyze and prevent fraudulent transactions not only with cryptocurrencies but also with other types of assets. In general, the prospects for using the model to predict fraudulent transactions in cryptocurrencies look promising and can lead to significant improvements in the security and efficiency of financial transactions.

Conclusion

An analysis of literature sources is performed and it is determined that the unresolved part of the problem, namely, the development of models, methods and technologies for data analysis to detect fraudulent transactions in the cryptocurrency market, including logistic regression, is still unresolved. A SWOT-analysis of security threats in the field of cryptocurrencies is carried out, their strengths and weaknesses are identified. To detect fraudulent operations, it is proposed to use technologies based on data analysis - logistic regression models. The algorithm for classifying fraudulent transactions is proposed using steps such as data preparation, sampling, algorithm application, accuracy evaluation, and model optimization. The R language was chosen to build the model, including its additional libraries, in the integrated environment of R Studio. A logistic regression model with 98 % accuracy that can check new transactions for fraud based on the Ethereum cryptocurrency has been developed. A model for predicting fraudulent transactions in cryptocurrencies can be of great importance to financial institutions and cryptocurrency exchanges, which can use it to prevent financial losses and improve transaction security. Prospects for further research are to improve the model to take into account new types of fraudulent schemes and increase its accuracy.

REFERENCES

Aponte-Novoa F.A., Povedano Álvarez D., Villanueva-Polanco R., Sandoval Orozco A.L., García Villalba L.J. (2022). On Detecting Cryptojacking on Websites: Revisiting the Use of Classifiers. Sensors. 2022; 22(23):9219. https://doi.org/10.3390/s22239219

M. Bartoletti, S. Lande, A. Loddo, L. Pompianu and S. Serusi (2021). "Cryptocurrency Scams: Analysis and Perspectives," in IEEE Access. Vol. 9. Pp. 148353–148373, 2021, doi: 10.1109/ACCESS.2021.3123894.

Ethereum Fraud Detection Dataset [Electronic resource] URL: https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset

Jiashi Feng, Huan Xu, Shie Mannor, Shuicheng Yan (2014). Robust Logistic Regression and Classification, Advances in Neural Information Processing Systems 27 (NIPS 2014)

Ismail Alarab, Simant Prakoonwit (2022). Effect of data resampling on feature importance in imbalanced blockchain data: comparison studies of resampling techniques, Data Science and Management. Volume 5. Issue 2. 2022, Pp. 66–76. ISSN 2666–7649, https://doi.org/10.1016/j.dsm.2022.04.003.

Kutera M. (2022). Cryptocurrencies as a subject of financial fraud. Journal of Entrepreneurship, Management, and Innovation, 18(4). 45–77. https://doi. org/10.7341/20221842

Logistic Regression [Electronic resource] URL: https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html

Mohammed, Nasser Hussain and Maram, Sai Charan Reddy, Fraud Detection of Credit Card Using Logistic Regression (March 15, 2022). Available at SSRN: https://ssrn.com/abstract=4135514 or http://dx.doi.org/10.2139/ssrn.4135514

F. Sabry, W. Labda, A. Erbad and Q. Malluhi (2020). "Cryptocurrencies and Artificial Intelligence: Challenges and Opportunities," in IEEE Access. Vol. 8. Pp. 175840–175858. 2020, doi: 10.1109/ACCESS.2020.3025211.

T. Wang and Y. Zhao (2022). "Credit Card Fraud Detection using Logistic Regression," 2022 International Conference on Big Data, Information and Computer Network (BDICN), Sanya, China, 2022. Pp. 301-305, doi: 10.1109/BDICN55575.2022.00064.

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

ОТВЕТСТВЕННЫЙ РЕДАКТОР
Мрзабаева Раушан Жалиевна

КОМПЬЮТЕРНАЯ ВЕРСТКА
Жадыранова Гульнур Даутбековна

Подписано в печать 15.09.2023.
Формат 60x881/8. Бумага офсетная. Печать - ризограф.6,5 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59