Data on global financial statistics demonstrate that total losses from fraudulent transactions around the world are constantly growing. The issue of payment fraud will be exacerbated by the digitalization of economic relations, in particular the introduction by banks of the concept of "Bank-as-a-Service", which will increase the burden on payment services.

The aim of this study is to synthesize effective models for detecting fraud in digital payment systems using automated machine learning and Big Data analysis algorithms.

Approaches to expanding the information base to detect fraudulent transactions have been proposed and systematized. The choice of performance metrics for building and comparing models has been substantiated.

The use of automatic machine learning algorithms has been proposed to resolve the issue, which makes it possible in a short time to go through a large number of variants of models, their ensembles, and input data sets. As a result, our experiments allowed us to obtain the quality of classification based on the AUC metric at the level of 0.977–0.982. This exceeds the effectiveness of the classifiers developed by traditional methods, even as the time spent on the synthesis of the models is much less and measured in hours. The models' ensemble has made it possible to detect up to 85.7 % of fraudulent transactions in the sample. The accuracy of fraud detection is also high (79–85 %).

The results of our study confirm the effectiveness of using automatic machine learning algorithms to synthesize fraud detection models in digital payment systems. In this case, efficiency is manifested not only by the resulting classifiers' quality but also by the reduction in the cost of their development, as well as by the high potential of interpretability. Implementing the study results could enable financial institutions to reduce the financial and temporal costs of developing and updating active systems against payment fraud, as well as improve the effectiveness of monitoring financial transactions

Keywords: digital payments, machine learning, automated synthesis, fraud detection, data science

UDC 336.717:343.3/.7
DOI: 10.15587/1729-4061.2020.212830

AUTOMATIC MACHINE LEARNING ALGORITHMS FOR FRAUD DETECTION IN DIGITAL PAYMENT SYSTEMS

O. Kolodiziev
Doctor of Economic Sciences, Professor,
Head of Department*
E-mail: kolodizev107@ukr.net
A. Mints
Doctor of Economic Sciences, Associate Professor,
Head of Department**
E-mail: mints_a_y@pstu.edu
P. Sidelov
Postgraduate Student**
E-mail: pavlo@sidelov.com
I. Pleskun
Postgraduate Student*
E-mail: inna.pleskun1993@gmail.com
O. Lozynska
Postgraduate Student*
E-mail: olgalozinskaya1@gmail.com
*Department of Banking and Financial Services
Simon Kuznets Kharkiv National University of Economics
Nauky ave., 9-A, Kharkiv, Ukraine, 61166
**Department of Finance and Banking
Pryazovskyi State Technical University
Universitetska str., 7, Mariupol, Ukraine, 87555

Received date 03.07.2020
Accepted date 22.09.2020
Published date 27.10.2020

1. Introduction

Data on the global financial statistics show that total losses from fraudulent transactions around the world are constantly growing. Although banks understandably do not advertise their losses, the urgency of the issue is shown by independent assessments from reputable news agencies. Thus, according to research [1, 2], from 2012 to 2017, the total losses from fraudulent bank card transactions (including credit, debit, and Prepaid Cards) worldwide more than doubled, from USD 11.27 billion to USD 22.8 billion. According to forecasts, by 2023, the same source assumes an increase in the volume of fraudulent transactions to USD 35.67 billion. Hereinafter, we shall consider such transactions to be fraudulent that involve funds in the client's account used by third parties, without the consent or permission of the account holder.

It should also be taken into consideration that there is now a trend around the world to introduce strategies to digitalize economic relations, in particular, the introduction of the "Bank-as-a-Service" concept by banks, which implies a significant increase in the burden on payment services. This can be expected to further exacerbate the issue of payment fraud.

For a long time, only passive methods were used to protect against payment fraud, whose essence is to make it difficult to access the customer's account unauthorized. At the end of the 20th century, however, the first active defense systems began to emerge. Initially, they were based on models for identifying fraudulent transactions formulated by experts in the form of verbal rules. For example, if credit card payments are made in different countries in a short period of time, the transaction is likely to be fraudulent. Such systems worked but allowed a significant number of false positives. In addition, to maintain the effectiveness of systems at a sufficient level, it is necessary to constantly update the models of fraud transaction detection as their effectiveness decreases over time.

The digital nature of payment relationships has made it possible to significantly automate the development of models for identifying fraudulent transactions through the use of machine learning and Big Data analysis techniques, as shown by studies [3, 4], and others. However, even in this case, the development of a system to detect fraudulent transactions and its maintenance require considerable time, as well as attracting qualified professionals. Taken together, this significantly increases the cost of such projects and makes them affordable only to large customers.

For these reasons, the next step in the development of anti-fraud transaction systems should be the transition to the automatic synthesis of models for identifying them. That would reduce the financial and time costs to implement financial monitoring projects. In addition, temporal cost decrease can be expected to improve the relevance of fraud detection models and, accordingly, increase their effectiveness in detecting new types of fraud.

Tools that implement automatic machine learning algorithms that are suitable for big data analysis have become available relatively recently. Therefore, the issues relating to their use for the synthesis of fraud detection models in digital payment systems, as well as the effectiveness of such models, are relevant. Of additional interest is the assessment of the cost of model synthesis and training, as well as the analysis of the ability to interpret machine learning results.

2. Literature review and problem statement

The synthesis of models for identifying fraudulent transactions in digital payment systems is a complex task. This requires addressing a series of problems related to the collection and preparation of data, choosing how to process them, interpreting the results, and analyzing their effectiveness [6]. Some aspects of this task have been considered in a series of studies, the most relevant of which are listed below.

The task of pre-processing transaction data in digital payment systems and the preparation of a training sample for further analysis by machine learning methods is tackled, specifically, in papers [3, 4, 7, 8]. Their analysis shows that, as part of the task under consideration, pre-processing of data has a series of features, such as unbalanced data classes and a small number of initial variables. This necessitates the use of techniques that improve the quality of input sampling data by including additional metrics. However, the set of such indicators cannot be clearly defined as it depends significantly on the conditions of operation of a particular payment system.

Much of the research has been done to examine the effectiveness of different models and methods for classifying transactions. For example, work [9] examines the use of a combination of Unsupervised and Supervised training methods. Article [10] suggested that a graph-based semi-supervised system should be used to solve the problem. This gradually expands the methodology for building fraud-detection systems; it should be noted that recently there has been more and more research on the use of resource-intensive technologies such as deep learning and artificial neural networks [11, 12]. The main issue with these methods is the poor interpretability of the models obtained as a result of their application whereas it is the identification of fraud factors that contributes to the effective control of it.

The analysis [7, 15] revealed that the authors most often imply the efficiency to be the quality of classification of fraudulent transactions. And on the same samples of data, the use of different methods makes it possible to obtain generally comparable results. At the same time, insufficient attention is paid to the accompanying efficiency factors, namely the costs of various resources (temporal, human, financial, computational). It is particularly important to reduce the time spent on the development and maintenance of systems for identifying fraudulent transactions, a factor that only [13] pays attention to.

A large part of the studies analyzed [3, 4, 7–11] refer, in one way or another, to a global open data analysis project in digital banking payment systems called "Fraud Detection with Machine Learning." The project has been carried out since 2013 at a web platform by researchers from Europe, the United States, and other countries of the world [14]. The results from the participants cover the main stages of the process of developing systems for identifying fraudulent transactions. The original data are available for free download and study, which has led to their use in this work. A big advantage of using the data of this project is the ability to compare the results obtained by different researchers.

Papers [3, 4, 7, 12, 16–18] focused on the use of such processes and methods of the input data analysis, which are fully under the control of the researcher. This is useful for investigating global patterns in data, but, in practical implementation, there are drawbacks such as high requirements for the researcher's qualifications, as well as a lot of time spent developing the analytical system and its subsequent maintenance.

To address these shortcomings, some authors have already attempted to partially automate the synthesis of machine learning models to analyze payment transactions [11, 13]. However, even though their results show the prospect of automatic synthesis of models within the subject area under consideration, this task has not yet been fully solved.

Thus, our analysis of the scientific literature has revealed that the task of identifying fraudulent transactions in digital payment systems requires a further solution. In particular, it is necessary to elucidate the issues related to improving the efficiency of both the models themselves and the processes of their synthesis and use.

3. The aim and objectives of the study

The aim of this study is to synthesize effective models for detecting fraud in digital payment systems using automatic machine learning algorithms.

To accomplish the aim, the following tasks have been set:

– to define the methodological basis of the study, which implies setting a task, justifying the choice of methods and algorithms for solving it, as well as the criteria for evaluating its effectiveness;

– to consider the construction of an information base of transactions in digital payment systems;

– to prepare and conduct experiments on the synthesis of fraud detection models in digital payment systems using automatic machine learning algorithms and to analyze their results;

– to analyze the possibilities of interpreting the resulting models and extracting knowledge from them about the most significant factors and patterns inherent in fraudulent transactions.

4. Materials and methods to study the fraudulent transactions in digital payment systems

General provisions.

The task of identifying fraudulent transactions in digital payment systems can be stated as follows:

Let there be an A dataset array containing a set of vectors $\overline{a}_i = \{a_{i1}, a_{i2}, ..., a_{i,m}\}$, each of which describes some of the parameters of a single user transaction in the payment system. The A array is formed and supplemented dynamically in the process of payment system operation. It should be noted that it includes only transactions that have already been confirmed by passive control methods (checking the balance in the account, checking the PIN, etc.).

However, these data are not enough to build a system to detect fraudulent transactions as the vector $\overline{a}_i$ lacks information about the nature of the operation, which can only be added *post factum*. This information is added to the A array as a binary parameter $f_i \in \{\text{true}|\text{false}\}$ after the final transaction calculations have been successfully completed, or after the claim from the account holder has been made for unauthorized write-offs. In statistics, the "true" or "positive" value typically corresponds to the onset of the event under study, that is, in our case, the transaction falls into the "fraudulent" category. The value of "false" or "negative" corresponds to a normal (genuine) transaction. The process of forming information about the transaction is conditionally shown in Fig. 1.

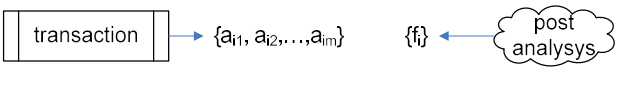

Fig. 1. Building an information base for a system for identifying fraudulent transactions

The result of the process shown in Fig. 1 is a database containing information about bank customer transactions made during the analyzed period and marking transactions to "regular" and "fraudulent."

The task of identifying fraudulent transactions is based on the assumption that there is some correlation $\phi(\cdot)$ between the values $\{a_{i1}, a_{i2}, ..., a_{i,m}\}$ and the corresponding f_i value, the knowledge of which would make it possible to determine the value of f_i directly, based on the values $\{a_{i1}, a_{i2}, ..., a_{i,m}\}$ (Fig. 2).

In a given form (Fig. 2), the task under consideration is a binary classification problem and belongs to the group of predictive problems of data mining [19]. Solving this type of problem employs a "Supervised learning" approach. It implies the division of the input sample of data into a training and test sample, after which the parameters of the classifiers are set up based on the training sample and controlled based on the test sample.

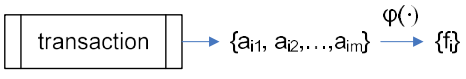

Fig. 2. The process to identify fraudulent transactions based on established dependences

Choosing methods to solve the problem of identifying fraudulent transactions.

There are various methods to solve the problems of binary classification. In the context of this study, it is advisable to divide them into "fast" and "slow." Among the "fast" methods that require relatively little computational cost, as regards the problem of fraud detection, one can distinguish discriminatory analysis [16], decision trees [20], random forests [21], Bayesian networks [17], and others. "Slow" methods that require high computational costs but give the best result of classification include artificial neural networks. And both the neural networks of classical architecture [18] and the more complex convolutional neural networks are successfully used [22].

The disadvantage of most of these methods is poor interpretability. That is, the resulting model is a kind of "black box" whose principle of action cannot be analyzed.

Exceptions include decision trees and methods built on their basis. Decision trees themselves and the methods to build them are a relatively "weak" classifier, that is, they have less accuracy compared to "slow" methods. However, this issue can be resolved by scaling, as it has been proven that a strong set of "weak" classifiers can be used to assemble a strong one [24]. The methods of synthesis of a strong classifier from the weak ones are termed "boosting". As a result, one can get an ensemble of models, which has both high accuracy and good interpretability.

At the same time, the task of synthesis of an ensemble of models has the attributes of a combinatorial one. Therefore, its solution requires significant computational resources. The synthesis algorithms of tree-based classification models are constantly improving. Currently, the best results are shown by the algorithms XGBoost and LightGBM. The first one is a little more accurate in some cases but the second one works almost twice as fast [21, 29].

The use of model ensembles in the task of detecting fraudulent transactions is not yet well understood but the available evidence suggests that the use of ensemble learning could produce good results [23]. Combined with the high potential for the interpretability of synthesized models, this has led to the choice of a given method in this study.

Choosing the performance metrics for transaction classification.

Since it is impossible to estimate the effectiveness of a particular model or their ensemble in advance, then, regardless of the technique to synthesize the models, the choice of the best classifier of several is made. Naturally, the ideology of sorting the solutions under automatic and manual ways is radically different. The main engine for obtaining the best solution in manual sorting is the experience and intuition of the developer, and with automatic – the algorithms and computing resources.

To measure the effectiveness of binary classification, the most common are the metrics based on the confusion matrix, whose general form is shown in Table 1.

Table 1

Confusion matrix

		Transaction genuine class	
		P	N
Predicted transaction class	P	TP	FP
	N	FN	TN

Abbreviations in the Confusion matrix are read and understood as follows:

– N – negative (genuine transaction);

– P – positive (fraud transaction);

– TN – true negative (genuine transaction, which is predicted as genuine);

– FN – false negative (fraud transaction, which is predicted as genuine);

– FP – false positive (genuine transaction, which is predicted as fraud);

– TP – true positive (fraud transaction, which is predicted as fraud).

Based on data from a confusion matrix, more than 15 different metrics are calculated and used in the statistical analysis of binary classifiers [16, 17]. In mathematics, the first- and second-class metrics are traditionally used to analyze the quality of binary classifiers. However, in highly unbalanced samples (such as payment transaction data), the results of calculating these errors become inconvenient to use. In such cases, precision and recall metrics are more adequate:

$$Precision = \frac{TP}{TP + FP}. \quad (1)$$

Precision shows how many of those who received the "fraud" label are indeed fraudulent.

$$Recall = \frac{TP}{TP + FN}. \quad (2)$$

Recall shows how many "fraud" transactions have been identified from the total number of transactions. This metric is also called *Sensitivity*. Similarly, another common metric, *True Positive Rate* (TPR), is calculated.

One of the features of calculating the *Precision* and *Recall* metrics is that an improvement in one of them can usually be achieved by degrading the other. Thus, setting up a classifier comes down to finding the optimal balance between *Precision* and *Sensitivity*. It should be noted that errors associated with misclassification of positive outcomes tend to be much more costly than misclassification of negative outcomes. Therefore, when setting up classifiers, one tends to get a high value of *Recall* at an acceptable *Precision*.

Another common metric for evaluating binary classifiers is the graphic curve *Receiving Operating Characteristic* (ROC). The chart shows the ratio between *True Positive Rate* (TPR) and *False Positive Rate* (FPR) metrics when the classification threshold is changed.

$$FPR = \frac{FP}{FP + TN}. \quad (3)$$

The diagonal line on the ROC curve corresponds to a "useless" classifier, that is, a model based on random event classification. In practice, such models, of course, do not apply, but, in the analysis of binary classifiers, they are used as a "base" to determine their effectiveness. The further the chart moves away from the diagonal up, the better the quality of the resulting classifier.

In addition to the visual comparison, the *Area Under Curve* (AUC) metric is also calculated based on a ROC curve. It is obvious that the "useless" model corresponds to the value of AUC=0.5, and, for classifiers of varying degrees of efficiency, $AUC \in (0.5; 1]$.

There are also other metrics that make it possible to reveal in more detail an aspect of the effectiveness of the examined classifier [25]. However, there is a high level of correlation between the results obtained using different metrics, that is, if the solution is effective for the main metrics, it will be quite effective for others, as well.

5. Results from the use of automatic machine learning algorithms to detect fraud in digital payment systems

5. 1. Features of building an information base to detect fraudulent transactions.

As shown above (Fig. 1, 2) the process of identifying fraudulent transactions implies the existence of some information base containing data on customer transactions and their markup to "regular" and "fraudulent".

Raw data from the information system of banks, or payment organizations, contain a technical minimum of information about transactions, which includes a relatively small set of parameters, including:

– date and time;

– the amount of the transaction (in the transaction currency and the currency of the account);

– the currency of the transaction;

– customer ID;

– the merchant/terminal ID through which the transaction was made;

– the type of identification procedure carried out;

– the authorization code, or the issuer's bank response if the transaction is rejected.

For a series of reasons that are detailed below, this set of parameters is not sufficient to effectively identify fraudulent transactions. This necessitates additional data preparation procedures.

Thus, when using machine learning methods, additional information, which is usually contained in related relational databases, should be immediately added to the input sample. This is because machine learning methods are focused on analyzing data arranged in flat tables.

In order to effectively detect fraud, in addition to the information available in the databases of the bank or payment organization, it is necessary to add to the data external information (for example, *a posteriori* assessment of various risks associated with the transaction) [26, 27].

The process of adding such information is termed "feature augmentation." Its necessity is due to the fundamental limitations inherent in methods focused on line-by-line data analysis, which were formulated for the first time in [28] for the perceptron neural networks, but are also true for most other machine learning methods:

– the inability to generalize their characteristics to new stimuli or new situations;

– the inability to analyze complex situations in an external environment by decomposing them into simpler ones;

– the limitations in problems related to the invariant representation of images.

Expanding the feature augmentation by adding information that clarifies the current situation reduces the impact of the specified restrictions. This information can be of several kinds:

– additional information that is not contained in the transaction's original data, but expands knowledge about it (for example, associated risks);

– the information that can be obtained as a result of vertical data analysis (for example, the average, maximum, minimum parameter values);

– the information obtained from empirical analysis models, which are usually formulated as a "condition" – "consequence" (for example, signs of transactions subject to mandatory financial monitoring). In fact, such models can be considered micro-expert systems.

An analysis of research in the field of monitoring and detection of fraudulent transactions [4, 22] has made it possible to formulate the following list of additional parameters for building a machine learning information base:

– the assessment of the risk of the terminal/merchant in which the transaction takes place;

– the assessment of the merchant's risk in which the previous transaction was made;

– the assessment of the risk of the merchant category;

– the estimates of geographic risk (continental, country, regional);

– assessing the risk of the card issuer;

– assessments of other risks (related to the age and gender of the client, language group, place of the previous transaction, transaction amount, etc.);

– the total amount of customer transactions for the period under review;

– the minimum amount of a customer's transaction for the period under review;

– customer data (age, gender, etc.);

– additional transaction parameters (the use of special identification technologies, time, information about accepting, or rejecting the transaction, etc.).

Thus, the original dataset can be expanded from 8 parameters to 25–30 (depending on the availability of information in the bank's databases and the available risk assessment capabilities).

In practice, building a representative information base for investigating fraudulent transactions involves solving the problem of finding a large enough sample of reliable evidence for analysis. The issue is due to that employees of banks and payment organizations consider the transaction data of their customers as confidential, and provide them for research only to commercial developers, subject to contractual terms.

The only dataset with real-world data containing parameters similar to the above is available for free study as part of the Big Data Mining and Fraud Detection Project [14]. The dataset contains actual credit card payments from customers of Western European banks, and has the following parameters:

– the dataset contains 284,807 transactions, including 492 cases of fraud identified *post-factum* (Fig. 1). This represents only 0.172 % of the total sample. Thus, the data sample is heavily unbalanced. This skew is typical of real data of digital payment systems and is due to the fact that the actual number of fraudulent transactions is relatively small;

– the sample contains 30 inputs and 1 output variable. Of these, only 3 variables ("Time," "Amount," "Class") contain name-appropriate transaction data that are not subject to additional conversions. The remaining variables, demanded by the European law, for the preservation of confidentiality, are converted into dimensionless values, and their titles are replaced with conditional ($V1$, $V2$, ..., $V28$);

– there are no missed values in the dataset;

– there are no cardholder identifiers, so all transactions can be considered independent of each other.

The list of parameters contained in this dataset is listed in [7]. This set generally corresponds to the above list of parameters, which makes it possible to use it in our study. It should also be noted that in accordance with the 2016 ruling of the European Union (General Data Protection Regulation, GDPR), personal data can be provided only in a fully anonymized form, which does not allow their deanonymization. Therefore, the dataset used is processed using the main component method; most of the variable names have been replaced with conditional ones. Despite some discomfort in interpreting the results, this can be interpreted as positive as part of the study's goal, as it reduces the subjectivity of assessments.

5. 2. Description of experiments on the automatic synthesis of models of detection of fraudulent transactions

Choosing tools.

There are currently two main approaches to creating automated machine learning software:

1. Cloud-based solutions. They include data analysis proposals from leading software corporations – Amazon, Google, IBM, Oracle, Microsoft. All of them are commercial and provide a well-developed toolkit for dealing with customer data uploaded to the provider's cloud service. However, the National Bank of Ukraine, like many other regulators, requires banks and other subjects of primary financial monitoring to keep their data in Ukraine. This makes it impossible to use any cloud solution [5].

2. Using data automatically locally. There are both full-fledged commercial solutions and software products that are distributed free of charge under open Apache Foundation licenses, which have worse service capabilities. There are also hybrid solutions, based on free tools packaged in the user's graphical interfaces. The goal of creating such products is to automate the implementation of basic algorithms and methods of machine learning.

Among the latter, the choice of software products to implement ensemble learning is limited to options such as the MATLAB 2020 implementation, the Python programming language library, several R programming language libraries, and the specialized H_2O Driverless AI software product. However, only the last of these products provides the possibility of fully automatic operation, at which the participation of the human researcher is limited to the stages of stating the task and analyzing the results [29, 30]. In addition, a given product is available for free for academic applications, which has led to choosing it for this study.

Preliminary data analysis.

The most resource-intensive stages of model synthesis using automatic machine learning algorithms are experimenting.

It is therefore advisable to analyze the main characteristics of data before the machine learning procedures are launched in order to determine their suitability for further research.

These characteristics include, first of all, the distribution of individual parameters, the value of the correlation between them, the presence of outliers in the data. This information helps identify some patterns in the data before the experiment begins, as well as to detect and, in some cases, eliminate errors that could negatively affect the results of subsequent experiments.

The results of the analysis of relationships in data are shown in Fig. 3 in the form of a network correlation graph.

Analysis of Fig. 3 shows that all correlations identified in the sample are quite weak (the network graph shows them in the same shade). Since there are no parameters with a strong correlation, there is no need to thin the data.

One can also note the correlation between the output of variable *Class* (denoting that the transaction belongs to "genuine" or "fraud") and the V14 input variable. This suggests a significant impact of this variable on the result. The chapter below will discuss the importance of variables further.

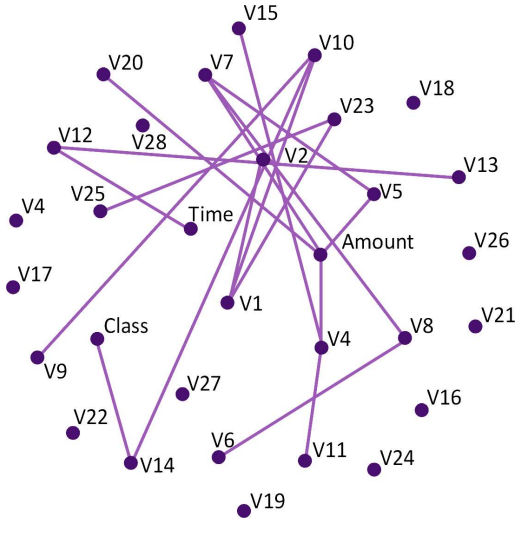

Fig. 3. Network correlation graph between variables in a credit card transaction dataset

At the same time, it should be taken into consideration that the correlation factor shows a linear degree of connection between the data elements. Therefore, at the non-linear nature of the links, including highly unbalanced samples, the results of correlation analysis should be taken critically. They can only be used to thin out parameters with strong mutual correlation.

Consider the emissions analysis using the example of "Transaction Amount" (Fig. 4).

Analysis of Fig. 4 shows that most transactions are characterized by amounts between EUR 0 and 2,500. Of the 284,807 transactions, only 4 were made for large sums. In developing a system for identifying suspicious and fraudulent transactions in the context of implementing a risk-oriented approach in the anti-money laundering process, this makes it possible to set a threshold for transactions through automatic or manual monitoring. Currently, the amount set by law, the equivalent of UAH 400,000, or approximately EUR 13,000 (at the time of writing the paper) [6], is used to determine the cut-off threshold for transactions subject to financial monitoring. In our case, its reduction to EUR 5,000 will not complicate the work of monitoring services but will reduce risks.

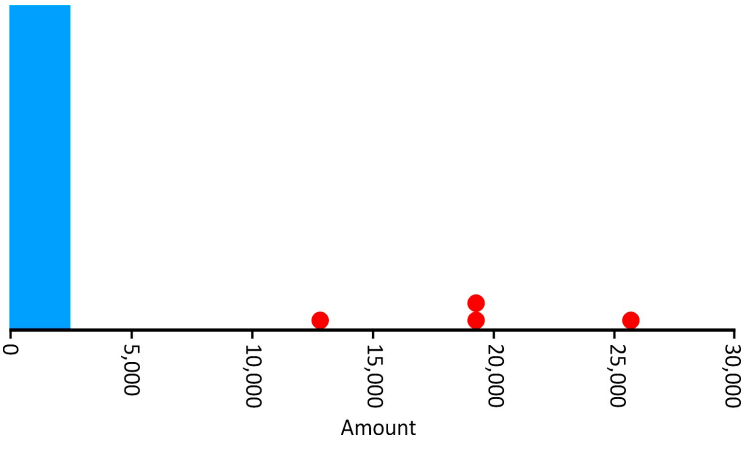

Fig. 4. Analysis of data emission based on the parameter "Transaction Amount"

Setting the parameters of the experiment.

In this case, parameters such as Accuracy, Time, Interpretability are set, determining, accordingly, the speed of the calculation of models, the necessary accuracy of the experiment, and the suitability of the results for verbal description.

Separately, in the process of setting the parameters of the experiment, one should highlight establishing a metric to determine the quality of classification of the sorted models. Using the selected automatic machine learning system limits the selection of criteria to the following: Accuracy, AUC, AUCPR, F05, F1, F2, Gini, LogLoss, MacroAUC, MCC (Matthews Correlation Coefficient). Features of the application of these criteria are given in Table 2.

Note that Table 2 misses the explicit forms of the above Precision and Recall metrics. Instead, comprehensive criteria F1, F0.5, F2, AUCPR, based on both metrics, can be used at the same time.

Our experiments have shown that in the context of operating a highly unbalanced sample of data, not all metrics were suitable for the procedure of automatic synthesis of machine learning models. Thus, when using the AUCPR metric based on the combination of Precision and Recall, the resulting model failed to adequately recognize fraudulent transactions. A similar situation occurred when using the MCC metric. At the same time, good results were obtained when using the LogLoss and Accuracy metrics; they are discussed in more detail below.

Thus, it can be concluded that the process of selecting the best metric in the task of detecting fraud in digital payment systems is weakly formalized. Several experiments with different criteria should be conducted to get the best result.

One should also distinguish between the metrics used in the process of automatic synthesis of models and metrics that are used to compare results. This study uses the AUC metric as the latter. Its value is given in many studies into our chosen data sample as part of the Big Data Mining and fraud detection project [14] and can be used to compare with the results by other researchers. The AUC metric was not used in the automatic model synthesis to keep the experiment virgin.

Table 2

Choosing quality classification metrics based on the characteristics of the task to be solved

Metric	Estimate based on	Feature of application
LogLoss	Probability	Assesses how close the projected model values are to the actual target value. Log Loss is a forecast uncertainty metric based on how different it is from the actual value
MCC	Class	Effective for using unbalanced data
F1	Class	Metric F1 is also called a balanced F-estimate or F-measure. F1 is useful if one needs to strike a balance between precision and recall
F0.5	Class	Analog to F1 but gives more weight to Precision
F2	Class	Analog to F1 but gives more weight to Recall
AUC	Class	Area Under the Curve of errors. One of the common metrics
AUCPR	Class	Area Under the Curve «Precision–Recall»: a convenient forecast quality metric for unbalanced data. A high score close to 1.00 shows that the classifier returns accurate results (high precision), as well as returns most of all positive results (high level of recall)
Accuracy	Class	The overall accuracy assessment improves the interpretability of the results

Description of our experiments and their main results.

We conducted several experiments on the automatic synthesis of the ensembles of models to identify fraudulent transactions in digital payment systems. The main differences between the conditions of the experiments were the use of different metrics for determining the quality of classification, as well as in the volumes of data analyzed.

Because the synthesis of an effective classifier employs the sorting out of a variety of variants by the genetic algorithm, the software requires significant computational resources. In our experiments, the *H2O Driverless AI* platform was launched under the Linux operating system on a hardware configuration that included 12 CPU cores and 32 Gb RAM.

Below are the results from two experiments to identify fraudulent transactions that yielded the best results.

Experiment-1 analyzed the full data set. The main classification quality metric for model synthesis is LogLoss. This classification experiment was completed in 3 hours and 37 minutes. The classifier construction's results used 20 of the 30 original features and 137 of the 3,872 constructed (synthetic) features.

The following steps have been taken in the process of automatically constructing the fraudulent transactions classifier:

– download data – identify column types;

– pre-process the objects – converting untreated objects into numerical ones.

– set up models and functions.

This stage combines the setting of random hyperparameters with the choice and generation of functions. Functions in each iteration are updated using the variable value from the previous iteration as a probabilistic one before deciding which new functions to create. The model with the best characteristics and functions is then transferred to the next stage. During Experiment-1, 21 models were trained and evaluated to assess the characteristics and parameters:

– evolutionary selection of features.

At this stage, a genetic algorithm is used to find the best set of model parameters and to transform the features to be used in the final model.

During Experiment-1, in this step:

– we found the best data representation for the final learning model by creating and evaluating 3,872 features in 33 iterations;

– 79 models have been trained and evaluated to further assess the features (synthetic parameters) that have been designed;

– the final model is an ensemble of 3 Imbalanced LightGBM models whose basic parameters are given in Table 3.

Table 3

The basic parameters of the ensemble of models obtained from Experiment-1

Model index	Type	Model weight	Cross-validation zone quantity	Input parameter quantity
0	Imbalanced-LightGBMModel	0.25	5	24
1	Imbalanced-LightGBMModel	0.4219	5	20
2	Imbalanced-LightGBMModel	0.3281	5	105

According to the report generated by the system, the stage of the evolutionary selection of features was the most resource-intensive. It accounted for 87 % of the system's total operating time.

Consider the quality indicators of the resulting transaction classifier.

The *Driverless AI* system automatically splits training data to determine the performance of the model settings and the stages of feature construction. For our experiment, *Driverless AI* randomly assigned two-thirds of the sample for training and 1/3 to test its results.

The confusion matrix of the experiment's results is given in Table 4.

Table 4

Confusion matrix for Experiment-1

		Transaction genuine class	
		P	N
Predicted transaction class	P	388	67
	N	104	284,248

It follows from the analysis of Table 4 that a given experiment analyzed 284,807 transactions, of which 492 were

fraudulent in reality. At the same time, the model correctly identified 388 (78.9 %) fraudulent transactions. The remaining 104 fraud transactions (21.1 %) were interpreted by the system as genuine. In addition, 67 genuine transactions were interpreted by the system as a fraud.

The Receiver Operating Characteristic Curve chart and the appropriate Area Under Curve (AUC) value are shown in Fig. 5.

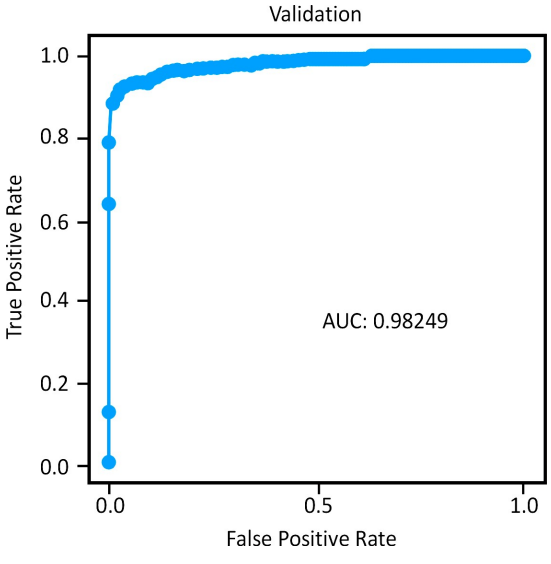

Fig. 5. ROC-curve for Experiment-1

The value of AUC=0.98249 is good in itself, but, when analyzing the unbalanced dataset, the final decision on the choice of the classifier can be made only after the AUC values are compared to all classifiers.

Experiment-2. The main differences from Experiment-1 were:

– accuracy is the main metric for the quality of classification in the synthesis of models.

– to reduce the calculation time, the input sample of data has been reduced to 100,000 lines.

The following results were obtained: an ensemble of the Imbalanced LightGBM models was built to predict the class, taking into consideration 30 original features from the input dataset. This experiment ended in 1 hour 8 minutes (1:08:00) using 3 out of 30 original features and 31 of the 2,803 constructed features. The total number of inputs is thus 34, which is almost 5 times less than that in Experiment-1.

The final model is an ensemble of 2 models of Imbalanced LightGBM whose basic parameters are given in Table 5.

Table 5

The basic parameters of the ensemble of models obtained from Experiment-1

Model index	Type	Model weight	Cross-validation zone quantity	Input parameter quantity
0	Imbalanced-LightGBMModel	0.6957	5	20
1	Imbalanced-LightGBMModel	0.3043	5	14
2	Imbalanced-LightGBMModel	0.0	5	20

It should be noted that in the parameters of the experiment it was indicated to use 3 models in the ensemble but, in the course of training, the automatic synthesis system determined that the best result is obtained with 2 models.

The confusion matrix of the experiment's results is given in Table 6.

Table 6

Confusion matrix for Experiment-2

		Transaction genuine class	
		P	N
Predicted transaction class	P	191	46
	N	32	99,730

It follows from the analysis of data in Table 6 that this experiment analyzed a total of 100,000 transactions, of which 223 were actually fraudulent. 191, or 85.6 % of fraudulent transactions, were correctly identified. The remaining 32 fraud transactions (14.4 %) were interpreted by the system as genuine. In addition, 46 genuine transactions were interpreted by the system as a fraud.

Fig. 6 shows the Receiver Operating Characteristic Curve chart and the corresponding Area Under Curve (AUC) value.

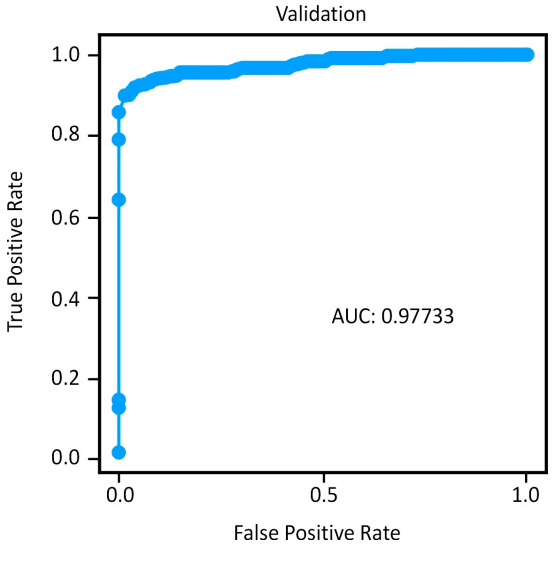

Fig. 6. ROC-curve for Experiment-2

The analysis of Fig. 6 shows that the AUC metric for Experiment-2 (0.97733) was slightly lower than that for Experiment-1 (0.98249) but this difference is negligible.

The comparative analysis of the effectiveness of the models' ensemble synthesized during Experiment-2 and Experiment-1, by calculating the *Precision* and *Recall* matrices, based on Table 3 and Table 4, produced the following results:

For Experiment-1:

$$Precision_1 = \frac{388}{388+67} = 0.85;$$

$$Recall_1 = \frac{388}{388+104} = 0.79.$$

For Experiment-2:

$$Precision_2 = \frac{191}{191+46} = 0.806;$$

$$Recall_2 = \frac{191}{191+32} = 0.857.$$

Thus, in the second experiment, we managed to get some better (by 8.5 %) *Recall* metric results showing how many "fraud" transactions were able to be identified from the total number of transactions.

The *Precision* metric, which shows how many transactions that received the "fraud" label were indeed fraudulent, was 5.1 % worse for the second experiment than that for the first.

A comparison of the results from experiments for statistical metrics, considered in Table 2, is given in Table 7.

Table 7

Comparative analysis of classification quality metrics in Experiments 1 and 2

Metric	Improvement direction	Value for a model selection criterion		Best
		Logloss (Experiment 1)	ACCURACY (Experiment 2)	
ACCURACY	higher	0.99945	0.99929	1
AUC	higher	0.98249	0.97733	1
AUCPR	higher	0.79529	0.76437	1
F0.5	higher	0.85889	0.84219	1
F1	higher	0.83447	0.84252	2
F2	higher	0.83601	0.8644	2
GINI	higher	0.96498	0.95466	1
LOGLOSS	lower	0.0031047	0.0040622	1
MACROAUC	higher	0.98249	0.97733	1
MCC	higher	0.8356	0.84357	2

It follows from Table 7 that, in terms of statistical classification quality criteria, in most cases, the best model was the model built during Experiment-1. At the same time, from the point of view of the bank, or payment organization, the result for the *Recall* metric is somewhat more important than the result for the *Precision* metric. Thus, in practice, the ensemble of models obtained as a result of Experiment-2 is likely to be selected.

Let us compare the results of our experiments with the results from other studies of the same Dataset, which were conducted by traditional methods, with the involvement of machine learning specialists. An analysis of the most extensive research in this area [7] shows that the various models developed by the author are characterized by an AUC value from 0.83 to 0.96. That is, according to formal criteria, the quality of the automatically synthesized model of the classifier was better than the result obtained by manual synthesis. Moreover, the cited work, published in 2015, actually summarizes the research of its author, which began in 2013 [4]. In other words, the development of these models took about 3 years.

Since the task of automatic synthesis of ensembles of models refers to combinatorial ones, it has great computational complexity. Consequently, the capability to effectively implement automatic machine learning algorithms is largely due to the available computing resources and their cost. To a large extent, the possibility of a positive result is due to the intensive development of computing, which reduces the cost of computing by half every 1.5–2 years. Therefore, in order to more accurately compare the effectiveness of different ways of synthesizing models, a correction should be made for the progress in the development of computing.

In 2020, for example, a productive enough hardware platform, including 12 CPU cores and 32 Gb RAM, was used for experiments. At the time of writing this paper, the computer's configuration used was priced at approximately USD 2,000. In this case, operating this configuration cannot be considered very comfortable because, to get the best result, one needs to conduct several experiments with different global settings, which takes several days.

If one takes advantage of the above-mentioned pattern, known as Moore's Law, we can determine that in 2013–2015 the cost of a hardware platform of comparable performance was USD 16,000–USD 32,000, or the time of experiments would have increased to about 1 month. Thus, even with the correction for the development of computing, the development of models for detecting payment fraud using automatic machine learning algorithms is significantly faster. Consequently, in the framework of the considered task, automatic learning makes it possible to find no less effective solutions than the traditional methods of model development.

5. 3. Analysis of the synthesized models from an interpretability perspective

Machine-synthesized models can be considered as some kind of a "black box", that is, a closed system that performs some function but is not known how. It is possible to increase its practical significance by revealing the operational principles of such a system, as well as the causal link between the input factors and an output variable. However, the interpretability of results is still one of the main problems relating to machine learning. To date, there are only methods that make it possible to assess the significance of individual factors for forecasting. And even such methods in themselves require a large cost of computing resources. This study, to determine the variables' significance, uses the Shapley calculation results' values:

$$\varphi_i(p) = \sum_{S \subseteq N/\{i\}} \frac{|S|!(n-|S|-1)!}{n!} \left(p\left(S \cup \{i\}\right) - p(S)\right), \lim_{x \to \infty}, \quad (4)$$

where $p\left(S \cup \{i\}\right)$ is the prediction of a model that uses the i factor; $p(S)$ is the prediction of a model that does not use the i factor; n is the total number of factors; S is the arbitrary set of factors without the i factor.

The result of analyzing the significance of variables, based on the results from Experiment-1, is given in Table 8. To reduce the size of the table, it provides only the top 10 variables in terms of significance.

Table 8 shows that of the original variables, only V14 made it to the top 10 in terms of significance but, at the same time, reached the top position. This is consistent with the results of the earlier correlation analysis (Fig. 3). The remaining variables in the top 10 are synthetic, that is, combined from others. However, because the composition of clusters of the synthetic variables is also given in Table 8, its analysis allows us to note the high significance of the V17, V10, V4, V18 variables, as well as some others.

Table 8

The significance of the input variables of the models' ensemble built as a result of Experiment-1

No.	Attribute	Description	Synthesis method	Relative importance
1	7_V14	V14 (Orig)	None	1.0
2	38_ClusterDist50: V10: V14: V17: V4.14	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 14	Cluster distance	0.6031
3	58_ClusterID50: V10: V14: V17: V18: V4	Assigning a centroid after segmenting the columns ['V10', 'V14', 'V17', 'V18', 'V4'] into 50 clusters	58_ClusterID50: V10: V14: V17: V18: V4	0.6018
4	38_ClusterDist50: V10: V14: V17: V4.15	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 15	Cluster distance	0.5321
5	66_ClusterDist20: V10: V14: V4.2	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V4'] into 11 clusters. Distance to cluster # 2	Cluster distance	0.4933
6	38_ClusterDist50: V10: V14: V17: V4.17	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 17	Cluster distance	0.4
7	38_ClusterDist50: V10: V14: V17: V4.13	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 13	Cluster distance	0.3635
8	37_ClusterDist10: V10: V14: V4.2	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V4'] into 7 clusters. Distance to cluster # 2	Cluster distance	0.325
9	38_ClusterDist50: V10: V14: V17: V4.5	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 5	Cluster distance	0.2668
10	38_ClusterDist50: V10: V14: V17: V4.8	Distance to the cluster center after parsing the columns ['V10', 'V14', 'V17', 'V4'] into 18 clusters. Distance to cluster # 8	Cluster distance	0.259

Because the resulting classifier is based on the decision trees, representing major dependences as a tree is one of the best ways to visualize and interpret the established dependences (Fig. 7).

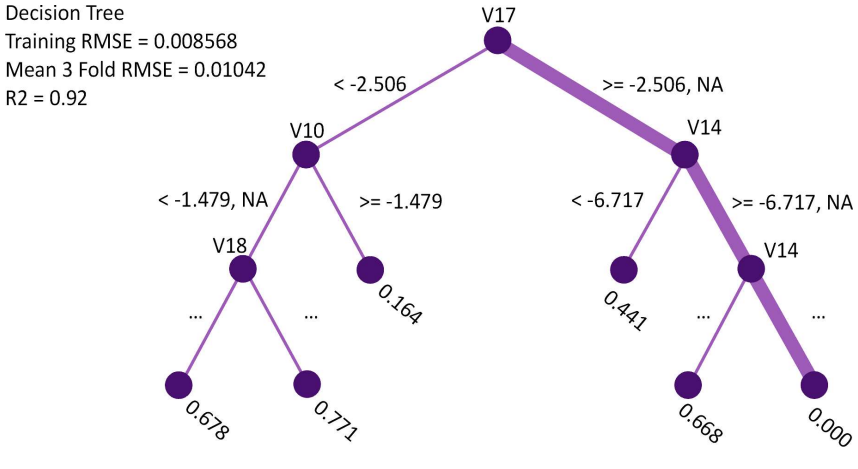

Fig. 7. Decision tree built in the process of interpreting the results from Experiment-1

An analysis of the decision tree, in this case, makes it possible to evaluate only the set of variables that have the strongest impact on the result. These include V17, V14, V10, V18, which is in good agreement with a set of variables obtained using Shapley values. For the case of impersonal variables, analyzing Fig. 7 does not make it possible to extract more information. However, if the input variables make economic sense, the decision tree becomes one of the best tools for machine learning in terms of interpretability of results.

6. Discussion of the application of automatic machine learning to synthesize payment fraud detection models

Our analysis of the basic parameters for the task of detecting fraudulent transactions has revealed that it is inherently a binary classification problem; therefore, an appropriate methodological base can be used to solve it. This finding has shaped the further course of our study, including the selection of a group of methods based on the principle of "supervised learning", as well as appropriate tools and metrics for evaluating the quality of the solution.

At the same time, given the very small number of fraudulent transactions in the total volume of payment system transactions (significantly less than 1 %), the task to detect them can also be considered as a search for anomalies in the data. In this case, the problem to be solved should be attributed to the tasks of the descriptive group. In solving such problems, an approach based on the principle of "unsupervised learning" [32] can be applied. In this case, the input

sample is divided into clusters, without taking into account the f_i feature (Fig. 1), after which the clusters containing the largest number of fraud values are determined, and the profiles of these clusters are interpreted as the attributes of potentially fraudulent transactions. It is advisable to check the effectiveness of this approach in further research.

The issues of pre-preparation of data and the formation of an information base for research are still not sufficiently formalized. Attempts to use raw transaction data in payment systems as the only source of information inevitably lead to inefficient solutions. This is predetermined by that in addition to internal factors, there are still a large number of external factors that also influence the likelihood of a given transaction to be fraudulent. Our paper offers the main directions of information search for expanding the input data vector and an indicative list of additional parameters. However, it remains to be studied whether our list of parameters is exhaustive and whether additional parameters can be added to it that could improve the quality of classification. In addition, the process of pre-preparation of data takes quite a long time, due to its weak formalization.

Our study has shown that in terms of the quality of classification (Fig. 5, 6) the model synthesized in an automatic mode is at least as good as the results obtained by other researchers using traditional approaches to data analysis. Formally, the resulting classifier has proven to be even somewhat more accurate. Thus, in our experiments, the matrices' values of AUC=0.97733 and AUC=0.98249 were obtained while other authors, when studying the same sample of data, received the values of AUC=0.96 and below. These results can be explained by that the synthesis of models sorts out many variations of both the composition of input parameters and the structure of the models. In addition, it is obvious that the problem itself is well suited to solving it by automatic machine learning methods.

The experiments also helped determine the difference between results obtained using the global optimality criteria, based on Logloss and Accuracy metrics. As shown by data from Table 7, using Logloss makes it possible to obtain some better results.

A significant advantage of the methods considered to detect fraudulent transactions is their good interpretability. An analysis of Table 8, automatically generated at the end of the models' ensemble synthesis process, makes it possible to establish the effect of different variables in the input vector of the data on the result. The possibility of obtaining such results is due to the use of models based on decision trees.

Among the drawbacks of using automatic machine learning algorithms, we, first of all, must note the cost of computation, which remains quite high, as noted above. Moreover, if the system's resources are not enough for the current needs of automatic machine learning algorithms, the process of model synthesis would be interrupted. As the data sample size increases, so does the cost of computing resources. In our experiments, the rate of such growth exceeded the linear one because the processing of a sample of 100,000 lines took 1 h 08 m while the processing of the sample of 284,000 lines took 3 hours 36 m.

Among the factors requiring further study is the lack of formalization of procedures for selecting the optimal classification quality metric for the automatic model synthesis algorithm operation. This now requires several experiments to select the best outcome. Therefore, the formalization of these procedures could significantly reduce the overall cost of time and computing resources.

In practical terms, the results reported here can be improved by using internal banking data that are free of the limitations of the dataset used. In particular, the considered observations cover only one month – September [14], that is, they do not take into consideration the seasonal features of credit card fraud.

Another factor influencing the results obtained is that the dataset has been devoid of cardholder identifiers, so all transactions are considered independent of each other. The inclusion of Card IDs in the dataset would make it possible to compile individual profiles based on customer transaction data and further use the identified deviations from these profiles as an additional indicator in fraud transaction detection.

These factors necessitate additional research during the implementation of our results in the operation of real banks and payment organizations. At the same time, there are areas to further improve the results of the study [31]. Specifically, it would be useful to develop a solution to the problem in question from the perspective of finding anomalies in the data and to consider its advantages and disadvantages, compared to the one proposed in our paper.

7. Conclusions

1. Underlying the formation of the methodological base of our study is the hypothesis that there is a correlation between the vector of transaction parameters and whether the transaction is fraudulent or not. Accordingly, the task of identifying fraudulent transactions in digital payment systems has been stated as a binary classification problem. Based on such a statement of the problem, the tools to solve it have been selected (the ensembles of classifiers based on the decision trees XGBoost and LightGBM). The criteria for assessing the effectiveness of the considered problem are based on an analysis of the results' confusion tables. Specifically, Precision and Recall metrics have been used.

2. Our experiments involved data on the actual credit card payments of customers in Western European banks. The dataset contains 284,807 observations, including 492 cases of fraud, representing 0.172 % of the total sample. Thus, the data sample is highly unbalanced, which is typical of the considered problem.

3. The result of our experiments, carried out by using automatic machine learning algorithms that make it possible, in a short time, to sort out a large number of variants of models and the composition of input data, is the quality of classification, in terms of the AUC metric, from 0.97733 to 0.98249. This is a high result and exceeds the effectiveness of the classifiers developed by traditional methods. The ensemble of models has allowed us to detect up to 85.7 % of fraudulent transactions in the sample. At the same time, the accuracy of detecting fraudulent transactions is also quite high (79–85 %).

4. The most significant variable is the risk inherent in the merchant's terminal. That is, in order to effectively execute financial monitoring, it is first of all necessary to block the identified channels of illegal transactions.

Acknowledgments

This paper is based on the results within the fundamental State-funded theme No. 54/2018-2020 "Risk-oriented approach in countering money laundering, financing of terrorism, and the proliferation of weapons of mass destruction".

References

1. The Nilson Report (2013). Issue 1023. Available at: https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1023
2. The Nilson Report (2017). Issue 1118. Available at: https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1118
3. Pozzolo, A. D., Caelen, O., Johnson, R. A., Bontempi, G. (2015). Calibrating Probability with Undersampling for Unbalanced Classification. 2015 IEEE Symposium Series on Computational Intelligence. doi: https://doi.org/10.1109/ssci.2015.33
4. Dal Pozzolo, A., Caelen, O., Waterschoot, S., Bontempi, G. (2013). Racing for Unbalanced Methods Selection. Lecture Notes in Computer Science, 24–31. doi: https://doi.org/10.1007/978-3-642-41278-3_4
5. Polozhennia pro orhanizatsiyu zakhodiv iz zabezpechennia informatsiynoi bezpeky v bankivskiy systemi Ukrainy 28.09.2017 No. 95. Available at: https://zakon.rada.gov.ua/laws/show/v0095500-17#Text
6. Pro zapobihannia ta protydiu lehalizatsiyi (vidmyvanniu) dokhodiv, oderzhanykh zlochynnym shliakhom, finansuvanniu teroryzmu ta finansuvanniu rozpovsiudzhennia zbroi masovoho znyshchennia 2020, No. 25, st. 17. Available at: https://zakon.rada.gov.ua/laws/show/361-20#n831
7. Dal Pozzolo, A. (2015). Adaptive Machine learning for credit card fraud detection. Université Libre de Bruxelles. Available at: http://di.ulb.ac.be/map/adalpozz/pdf/Dalpozzolo2015PhD.pdf
8. Russac, Y., Caelen, O., He-Guelton, L. (2018). Embeddings of Categorical Variables for Sequential Data in Fraud Context. Advances in Intelligent Systems and Computing, 542–552. doi: https://doi.org/10.1007/978-3-319-74690-6_53
9. Carcillo, F., Le Borgne, Y.-A., Caelen, O., Kessaci, Y., Oblé, F., Bontempi, G. (2019). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences. doi: https://doi.org/10.1016/j.ins.2019.05.042
10. Lebichot, B., Braun, F., Caelen, O., Saerens, M. (2016). A graph-based, semi-supervised, credit card fraud detection system. Complex Networks & Their Applications V, 721–733. doi: https://doi.org/10.1007/978-3-319-50901-3_57
11. Lebichot, B., Le Borgne, Y.-A., He-Guelton, L., Oblé, F., Bontempi, G. (2019). Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection. Recent Advances in Big Data and Deep Learning, 78–88. doi: https://doi.org/10.1007/978-3-030-16841-4_8
12. Georgieva, S., Markova, M., Pavlov, V. (2019). Using neural network for credit card fraud detection. Renewable energy sources and technologies. doi: https://doi.org/10.1063/1.5127478
13. Lucas, Y., Portier, P.-E., Laporte, L. et. al. (2019). Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs. Available at: https://www.researchgate.net/publication/335600419
14. Fraud detection with machine learning. Available at: https://www.researchgate.net/project/Fraud-detection-with-machine-learning
15. Wei, W., Li, J., Cao, L., Ou, Y., Chen, J. (2012). Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web, 16 (4), 449–475. doi: https://doi.org/10.1007/s11280-012-0178-0
16. Mahmoudi, N., Duman, E. (2015). Detecting credit card fraud by Modified Fisher Discriminant Analysis. Expert Systems with Applications, 42 (5), 2510–2516. doi: https://doi.org/10.1016/j.eswa.2014.10.037
17. Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., Cela-Díaz, F. (2010). Statistical Methods for Fighting Financial Crimes. Technometrics, 52 (1), 5–19. doi: https://doi.org/10.1198/tech.2010.07032
18. Patidar, R., Sharma, L. (2011). Credit card fraud detection using neural network. International Journal of Soft Computing and Engineering (IJSCE), 1, 32–38. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.301.8231&rep=rep1&type=pdf
19. Mints, A. (2017). Classification of tasks of data mining and data processing in the economy. Baltic Journal of Economic Studies, 3 (3), 47–52. doi: https://doi.org/10.30525/2256-0742/2017-3-3-47-52
20. Sahin, Y., Bulkan, S., Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. Expert Systems with Applications, 40 (15), 5916–5923. doi: https://doi.org/10.1016/j.eswa.2013.05.021
21. Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., Jiang, C. (2018). Random forest for credit card fraud detection. 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). doi: https://doi.org/10.1109/icnsc.2018.8361343
22. Fu, K., Cheng, D., Tu, Y., Zhang, L. (2016). Credit Card Fraud Detection Using Convolutional Neural Networks. Lecture Notes in Computer Science, 483–490. doi: https://doi.org/10.1007/978-3-319-46675-0_53
23. Zareapoor, M., Shamsolmoali, P. (2015). Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier. Procedia Computer Science, 48, 679–685. doi: https://doi.org/10.1016/j.procs.2015.04.201
24. Schapire, R. E. (1990). The strength of weak learnability. Machine Learning, 5 (2), 197–227. doi: https://doi.org/10.1007/bf00116037
25. Sammut, C., Webb, G. I. (Eds.) (2010). Encyclopedia of machine learning. Springer. doi: https://doi.org/10.1007/978-0-387-30164-8
26. Vnukova, N., Kavun, S., Kolodiziev, O., Achkasova, S., Hontar, D. (2019). Determining the level of bank connectivity for combating money laundering, terrorist financing and proliferation of weapons of mass destruction. Banks and Bank Systems, 14 (4), 42–54. doi: https://doi.org/10.21511/bbs.14(4).2019.05
27. Malyaretz, L., Dorokhov, O., Dorokhova, L. (2018). Method of Constructing the Fuzzy Regression Model of Bank Competitiveness. Journal of Central Banking Theory and Practice, 7 (2), 139–164. doi: https://doi.org/10.2478/jcbtp-2018-0016
28. Minsky, M., Papert, S. (2017). Perceptrons. MIT Press. doi: https://doi.org/10.7551/mitpress/11301.001.0001

29. Driverless AI Documentation - Overview. Available at: http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html
30. Driverless AI Documentation - Scorers. Available at: http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/scorers.html
31. Fabuš, M., Dubrovina, N., Guryanova, L., Chernova, N., Zyma, O. (2019). Strengthening financial decentralization: driver or risk factor for sustainable socio-economic development of territories? Entrepreneurship and Sustainability Issues, 7 (2), 875–890. doi: https://doi.org/10.9770/jesi.2019.7.2(6)
32. Mints, O., Marhasova, V., Hlukha, H., Kurok, R., Kolodizieva, T. (2019). Analysis of the stability factors of Ukrainian banks during the 2014–2017 systemic crisis using the Kohonen self-organizing neural networks. Banks and Bank Systems, 14 (3), 86–98. doi: https://doi.org/10.21511/bbs.14(3).2019.08

A method of estimating the effective data rate in channels of the Standard 802.11 was proposed. It provides for the measurement of the main energy parameter using the software and hardware of the subscriber device. This method is based on the empirical models of statistical relationships between the main parameters of the channel which are obtained on the basis of experimental studies using monitoring algorithms. The solutions obtained during the implementation of this method make it possible to take into account the maximum possible number of destabilizing factors and significantly reduce the time of assessment of the effective data rate. It should be noted that this method can be used for technical diagnostics of wireless networks of Standards 802.11x at the stages of network design and operation.

It was established that when using the coefficient of energy efficiency, a significant error in the displacement of the points of intersection of the linear and logarithmic mathematical model occurs. This can lead to a discrepancy between the mathematical estimates of the effective data rate and real values. The statistical relationship gives a smaller error; however, it increases requirements for empirical studies to obtain the maximum possible reliability.

One of the features of the proposed method is the reliability of assessment of the effective data rate. This reliability depends on three main factors: accuracy of assessing the results based on which the mathematical model was obtained; estimation of fluctuation intervals and characteristics of the Standard 802.11 equipment of different manufacturers. The last factor can be considered as a disadvantage that involves the creation of a database of parameters of the model of statistical relationship for different devices with correction coefficients

Keywords: wireless channel, Standard 802.11, effective data rate, signal strength, assessment method, statistical relationship

UDC 621.391.8
DOI: 10.15587/1729-4061.2020.213834

DEVELOPMENT OF A METHOD FOR ASSESSING THE EFFECTIVE INFORMATION TRANSFER RATE BASED ON AN EMPIRICAL MODEL OF STATISTICAL RELATIONSHIP BETWEEN BASIC PARAMETERS OF THE STANDARD 802.11 WIRELESS CHANNEL

D. Mykhalevskiy
PhD, Associate Professor
Department of Telecommunication Systems and Television
Vinnytsia National Technical University
Khmelnytske highway, 95, Vinnytsia, Ukraine, 21021
E-mail: adotq@ukr.net

Received date 21.08.2020
Accepted date 07.10.2020
Published date 22.10.2020

1. Introduction

At the present stage of technical development, wireless networks of the 802.11x standard family have become quite widespread [1]. Such networks are cost-effective, easy to build and maintain and have a fairly high bandwidth capacity. The range of their use is quite wide: from combining devices of the Internet of Things concept to providing access to television and infocommunication services [2].

The widespread use of wireless networks of Standard 802.11 leads to a series of negative factors causing delays and errors during traffic sessions. However, the main task set at the design stage is the achievement of the maximum possible bandwidth capacity of the channel and its stability for the