DETECTING CREDIT CARD FRAUD USING MACHINE LEARNING ALGORITHMS

Abstract. Today the banking sector offers its clients many different financial services such as ATM cards, Internet banking, Debit card, and Credit card, which allows attracting a large number of new customers. This article proposes an information system for detecting credit card fraud using a machine learning algorithm. Usually, credit cards are used by the customer around the clock, so the bank’s server can track all transactions using machine learning algorithms. It must find or predict fraud detection. The dataset contains characteristics for each transaction and fraudulent transactions need to be classified and detected. For these purposes, the work proposes the use of the Random Forest algorithm.

Keywords. Fraud detection, credit card, machine learning algorithm, Random Forest algorithm.

Introduction. Domestic banking institutions, like other economic entities, operating in conditions of unpredictability, uncertainty, threats and dangers. At the same time, the role of banks is constantly increasing. The banking system is an important element of the economy and has a significant, multifaceted impact on all
aspects of society. Banks perform a wide range of operations, such as: accumulation of funds and savings, credit settlement and other operations. In addition, almost all financial flows of individuals and legal entities pass through banks.

Modern banks operate under conditions of the destabilizing influence of both external and internal factors. Nowadays there is an acute issue of ensuring the economic security of the state banking system as the basis of the financial system.

The presence of the above problems is caused by a number of factors, among which, first of all, it should be noted the low level of economic security of banks, caused by the deficiencies in the functioning of the existing security management systems of the banking business, should ensure the implementation of the main interests, priority goals of banks, protection from the impact of negative factors.

One of the main components of the economic security system is the monitoring of banking operations as a form of countering fraud in various functional areas of the bank's activities.

Constant monitoring of fraudulent activities as one of the mandatory elements of the bank's economic security management system is, in our opinion, an important and urgent issue in the context of ensuring the stable and effective functioning of the domestic banking system.

**Literature review.** Many works of domestic and foreign scientists are devoted to the study of the issues of detecting fraud in banks carried out by bank personnel. The work [1] provides a classification of methods of fraud in the banking sector. In work [2] frauds carried out by the bank's personnel are considered in the context of distortion of the bank's financial statements.

The work [3] provides a description of fraudulent activities depending on the type of banking operation. In [4], the Data Mining technology (associative analysis) is considered, which can be used to detect fraud in banks, which is carried out by bank personnel. The work [5] provides examples of using artificial neural networks to detect personnel fraud on the basis of public financial statements.

The most common bank frauds are money laundering, credit fraud and asset misappropriation [6].
Financial fraud is characterized by the subject of misappropriation, location of financial fraud, sphere of existence of financial fraud, subjects of financial fraud, direction of financial fraud, manipulations that make financial fraud, and tools for its implementation [7].

The lion's share of banking fraud occurs with credit cards. Credit card fraud is known to involve illegal use of a credit card or its information without the knowledge of the owner. In work [8] it is noted that today, to detect such frauds, the following are widely used: logistic regression, which is able to solve categorical classification problems; support vector machine (SVM, Support Vector Machine), which is able to handle unbalanced data and complex relationships between variables; easy-to-use decision trees; self-organizing maps of Kohonen (SOM, Self-Organizing Map).

According to the authors of [9], in the presence of uncertainties, the best results are obtained by using fuzzy methods. Despite the quite decent results that the support vector machine gives, it is sensitive to an increase in the amount of data and cannot support large datasets [8].

In [10], a genetic algorithm is used for this, in which, instead of maximizing the number of correctly classified transactions, an objective function with variables showing the losses from misclassification is determined. Thus, the correct classification of some transactions is more important than others.

In [11], the Hidden Markov Model (HMM) is used to monitor the behavior of card account holders, which first learns the normal actions of the cardholder and then is used to detect fraudulent behavior. The work [12] uses the theory of fuzzy logic to monitor the behavior of card account holders.

**Proposed system and methods.**

Credit card fraud detection can be done using machine learning. Different Machine Learning approaches can be applied to this problem. In "Credit Card Fraud Detection: A case study" by Ayushi Agrawal, Shiv Kumar and Amit Kumar Mishra, combination of techniques is used like Genetic Algorithm, Behavior Based Technique and Hidden Markov Model. By this transaction is tested individually
and whatever suits the best is further proceeded. And the foremost goal is to detect fraud by filtering the above techniques to get better result. For more details see [1–5].

The Fraudulent Transaction Assessment workflow contains the following steps (Figure 1):

1. The client application sends information to the fraudulent transaction detection application.

2. The fraudulent transaction detection model determines the risk score (in the range 0-100) for the input data using a machine learning model that is trained using historical data. A score of 0 indicates that the forecast is considered to have the lowest possible risk, and a score of 100 indicates that the forecast is considered to have the highest possible risk.

3. If the risk score for a particular forecast falls below a predetermined threshold, no further action is taken.

4. If the risk assessment exceeds a predetermined threshold (for example, 90), the cycle starts automatically and sends forecasts for review by the bank staff. Bank staff review the transaction and make a decision (approve, reject, or send for further verification).

5. The result of approval or rejection is stored in the database. The data from the database can be used to retrain the fraudulent transaction detection model.

Fig. 1. Scheme of operation of the fraudulent transaction detection system
This paper proposes the use of the Random Forest algorithm (Fig. 2) to detect financial fraud in transactions of bank customers. The data on financial transactions of the bank's clients will be used as the initial data.

Random Forest is a combined classification method proposed by renowned scientists and includes a decision tree (DT) method is the main classifier and sets a large number of trees.

The data provided (Figure 3) contains financial transaction data as well as the target variable isFraud, which is the actual fraud status of the transaction, and isFlaggedFraud is an indicator that is used to indicate a transaction using some threshold value.

In the process of data preparation, the text values for the transaction type field are replaced with numeric values. It also removes fields that are not informative and necessary for building a machine learning model: nameOrig, nameDest, isFlaggedFraud. The data cleaning results are shown in Figure 4.
Since there are no “missing” and “garbage” values, there is no need for additional data cleaning, but data analysis is necessary, as the data contains huge variations in values in different columns. Normalization will also improve the overall accuracy of the machine learning model.

Let's analyze the amounts of transfers by type of transaction. The result of the analysis is presented in the form of a bar chart (Fig. 5).

As you can see from the graph, the largest volume of transactions is carried out in the form of transfers (TRANSFER), in second place are transactions related to cash withdrawal from the account (CASH_OUT) and in third place is the deposit of money to the account (CASH_IN).
The graph above shows that TRANSFER and CASH_OUT are also the only way to be scammed. Thus, we will focus on this type of transaction.

There are 2 fields that provide useful information for the fraud detection model: isFraud and isFlaggedFraud column. Based on the hypothesis, isFraud is an indicator that indicates “actual fraudulent transactions”, while isFlaggedFraud is an indicator that indicates that the system is preventing a transaction due to “certain thresholds” being triggered.

Having built a diagram of identified fraudulent transactions (Fig. 6), one can see that only 0.2% of such transactions are detected by the system.

![Percentage of identified fraudulent transactions](image)

**Fig. 6. Percentage of identified fraudulent transactions**

We will also build a graph of identified fraudulent transactions that were marked as fraudulent (Figure 7).

The graphs above clearly show the need for a software module that quickly and reliably detects fraudulent transactions.

The proposed software module uses the Random Forrest algorithm to classify a dataset by financial transactions.
The following steps form the foundation for any machine learning workflow:
1. Formulate the problem and determine the required data.
2. Receive data in an accessible format. Identify and correct missing data points / anomalies as needed.
3. Prepare data for the machine learning model.
4. Establish a baseline model to be investigated.
5. Train the model on training data.
6. Make predictions based on test data.
7. Compare the predictions with the known data from the test suite.

Since the first three stages have already been completed, let's move on to the stage of building a machine learning model.

There is one final step in preparing the data: splitting the data into training and test suites.

The following code splits datasets into test and training datasets.

```python
from sklearn.model_selection import train_test_split
train_X, test_X, train_y, test_y = train_test_split(X, y, test_size = 0.2, random_state = 121)
```

Next, using the Scikit-learn library and the random forest algorithm from skicit-learn, we create an instance of the model and train the model on the training data (Figure 9).
After training the model, you need to evaluate the quality of the model. To do this, it is necessary to determine the fraudulent transactions by test characteristics, after which it is necessary to compare the responses received by the model with the known responses. (Fig. 10).

The assessment showed that the model proposed in the work is able to identify fraudulent transactions with an accuracy of 77% (Fig. 11).

Thus, we can conclude that the random forest algorithm will perform better with a large amount of training data, but the speed will suffer during testing and application. More preprocessing techniques would also help.
Conclusion. The banking sector provides many services to clients, the credit card is one of the important services offered by major banks. Despite the fact that today there are a number of information systems and methods for detecting fraud, none of them can guarantee that all fraudulent activities will be detected. Thus, there is a need to develop a system for detecting fraudulent banking transactions that are carried out with a bank card. This paper proposes an approach to identify fraudulent banking transactions based on a machine learning algorithm - a random forest. In this work, a model is built and trained using a machine classification algorithm based on a training and testing dataset. The random forest algorithm showed a classification accuracy of fraudulent transactions of 77%.

References:

