ORIGINAL PAPER

DOI: 10.26794/2587-5671-2020-25-5-186-199
UDC 336.7(045)
JEL E58, G21, C53

Comparative Analysis of Machine Learning Methods to Identify Signs of Suspicious Transactions of Credit Institutions and Their Clients

Yu.M. Beketnova
Financial University, Moscow, Russia
https://orcid.org/0000-0002-1005-6265

ABSTRACT

In the field of financial monitoring, it is necessary to promptly obtain objective assessments of economic entities (in particular, credit institutions) for effective decision-making. Automation of the process of identifying unscrupulous credit institutions based on machine learning methods will allow regulatory authorities to quickly identify and suppress illegal activities. The **aim** of the research is to substantiate the possibilities of using machine learning methods and algorithms for the automatic identification of unscrupulous credit institutions. It is required to select a mathematical toolkit for analyzing data on credit institutions, which allows tracking the involvement of a bank in money laundering processes. The paper provides a comparative analysis of the results of processing data on the activities of credit institutions using classification **methods** — logistic regression, decision trees. The author applies support vector machine and neural network methods, Bayesian networks (Two-Class Bayes Point Machine), and anomaly search — an algorithm of a One-Class Support Vector Machine and a PCA-Based Anomaly Detection algorithm. The study presents the results of solving the problem of classifying credit institutions in terms of possible involvement in money laundering processes, the results of analyzing data on the activities of credit institutions by methods of detecting anomalies. A comparative analysis of the results obtained using various modern algorithms for the classification and search for anomalies is carried out. The author **concluded** that the PCA-Based Anomaly Detection algorithm showed more accurate results compared to the One-Class Support Vector Machine algorithm. Of the considered classification algorithms, the most accurate results were shown by the Two-Class Boosted Decision Tree (AdaBoost) algorithm. The research results can be used by the Bank of Russia and Rosfinmonitoring to automate the identification of unscrupulous credit institutions.

Keywords: suspicious transactions; money laundering; bank; credit institution; anomaly detection methods; machine learning

For citation: **Beketnova Yu.M. Comparative analysis of machine learning methods to identify signs of suspicious transactions of credit institutions and their clients. *Finance: Theory and Practice.* 2021;25(5):186-199. DOI: 10.26794/2587-5671-2021-25-5-186-199**

INTRODUCTION

Credit organizations [and other subjects of Federal Law No. 115-FZ dated 07.08.2001 "On Counteracting Legalization (Laundering) of Criminally Obtained Incomes and Financing of Terrorism"[1]] are on the first line in the fight against illegal financial activities. They are designed to collect and analyze information about their clients and their financial transactions and report in case of suspicious activity in Rosfinmonitoring. However, credit institutions can also be involved in illegal activities, deliberately covering up shadow schemes, or because of a weak internal control system.

Money laundering processes affect various sectors of the economy and are reflected, for example, in the amount of taxes and fees paid, money withdrawn abroad, the dynamics of the creation and liquidation of Russian legal entities, foreign trade, and the general state of crime, migration trends, etc.

Rosfinmonitoring is the central link in the Russian anti-money laundering system, interacting on this issue with other financial intelligence units, international organizations, federal executive authorities, law enforcement agencies, and credit institutions.

In the process of financial monitoring, the following negative manifestations can be identified on the part and/or in relation to credit institutions:

- using the banking infrastructure to organize schemes for the provision of shadow financial services and money laundering;
- withdrawal of funds from the bank — for example, the issuance of deliberately bad loans before bankruptcy or deliberate bankruptcy;
- bankruptcy of a credit institution.

In practice, one has to deal with a combination of these components at different degrees of their manifestation. Therefore, in the future, we will use the term "deviant component of a credit institution's activities", meaning one or more of the negative manifestations described above.

The features of the analysis of financial monitoring data and the identification of persons and organizations involved in money laundering are as follows:

1. Deviant subjects seek to hide their involvement in illegal activities and actively disguise themselves as law-abiding participants, using modern information technologies and expert knowledge of professionals in the field of finance, law, etc.
2. In connection with Clause 1, Rosfinmonitoring analysts need to analyze large volumes of heterogeneous data to identify hidden violations.
3. Carrying out this kind of analysis for each object requires from the analyst deep professional knowledge and practical experience in the subject area, on the one hand, and significant time resources, on the other.

When analyzing credit institutions, Rosfinmonitoring analysts have to operate with large amounts of information. We provide some statistics to illustrate this claim.

According to the official information of the Bank of Russia, as of 01.01.2020, there are 442 credit institutions and 618 branches[2] operating in the Russian Federation. Bank reporting contains hundreds of parameters. In addition, the federal database of Rosfinmonitoring maintains its own records of information on each credit institution — up to 50 data fields.

In addition to the initial data in the federal database of Rosfinmonitoring, statistics and additional identification information is generated for each type of object — addresses, data of identity documents, etc. Rosfinmonitoring's departmental data is

[1] Federal Law No. 115-FZ of 07.08.2001 "On Counteracting the Legalization (Laundering) of Criminally Obtained Incomes and Financing of Terrorism". Reference and legal system "Consultant Plus". URL: http://www.consultant.ru/document/cons_doc_LAW_32834/ (accessed on 04.10.2020).

[2] Reference book on credit institutions. URL: https://www.cbr.ru/banking_sector/credit (accessed on 14.11.2020).

enriched with information from various state registers, information on the foreign economic activity of entities, tax information.

The effectiveness of the anti-money laundering system largely depends on its ability to timely identify trends and patterns in the activities of entities, which necessitates the prompt receipt of objective assessments of the work of objects of financial monitoring.

Traditionally, state bodies use an approach to inspections of objects of supervision, which consists of the sequential assessment by an expert of one object of inspection after another. Such assessments may have expert subjectivity [1, 2]. In addition, this approach is resource- and time-consuming.

The growing volume of incoming information (approximately 20% annually) leads to a decrease in the efficiency of its processing. Those responsible for making decisions are forced to work with subjective analysis results for long periods of time for their receipt [3].

The analysis of the tasks of Rosfin-monitoring in combating money laundering showed that the real need for the number of analyzed objects is many times greater than the capabilities of analysts. This problematic situation requires prioritizing checks.

The diversity of information resources and a significant volume exclude the possibility of manual processing.

A transition is required from sequential expert inspections of individual objects to parallel massive automated inspections, considering modern methodological and instrumental capabilities in the context of the digital transformation of public administration.

Modern methods of data analysis and machine learning can serve as a necessary tool for this. And the automation of the process of identifying unscrupulous credit institutions based on machine learning methods will allow regulatory authorities to quickly identify and suppress illegal activities.

The problem of money laundering through the banking sector is now well understood. There are many publications devoted to risk-based banking supervision in order to combat money laundering and terrorist financing, for example [4–8]. Much of the research in this area is aimed at improving internal control rules in banks and introducing know-your-customer principles. At the same time, the deviant component of a credit institution's activities may be associated not only with the illegal activities of its clients but also with the actions of the bank's management. And studies devoted to automating the identification of unscrupulous banks to combat money laundering and terrorist financing are quite rare. This article is intended to fill this gap, since, unlike existing studies, it analyzes the issues of automating the identification of banks' involvement in money laundering.

Research hypothesis: the diagnosis of a bank's involvement in money laundering processes can be carried out using machine learning methods.

The aim of the study is to substantiate the possibility of using machine learning methods and algorithms for the automatic identification of unscrupulous credit institutions. To do this, it is necessary to select a mathematical toolkit for analyzing data on credit institutions, which makes it possible to diagnose the involvement of a bank in money laundering processes.

ANALYSIS OF WORKS OF DOMESTIC AND FOREIGN SCIENTISTS ON THE APPLICATION OF MACHINE LEARNING METHODS

The problem of identifying deviant objects of financial monitoring — shell companies, banks acting as platforms for money laundering, credit organizations on the verge of bankruptcy and withdrawal of financial assets, unscrupulous participants in the securities market — can be approached as a classification problem.

Let us turn to the methods and algorithms that allow us to solve similar problems in the field of financial monitoring.

Logistic regression

Logistic regression is a well-studied and widely used technique in statistics. In modern studies, highlighted in publications [9–11], logistic regression is used in combination with other methods or for comparison with them [12, 13].

Decision trees

Decision trees have a number of advantages when solving classification problems:

- efficient in computing and using computer memory, which makes them suitable for working with large amounts of data;
- the choice of functions is integrated into the learning and classification processes;
- non-parametric models, which allows processing data with different distributions.

One of the significant tasks in solving classification problems is the imbalance of classes. This research is devoted to its solution [14–18].

A Two-Class Decision Forest is a concrete implementation of a decision forest that works by building multiple decision trees and then voting for the most popular output class. Voting is one of the most well-known methods of obtaining results in the ensemble model. Trees with high prediction confidence will have more weight in the final ensemble solution. Decision forest has been successfully used in socio-economic research. The study [19], which deals with the issues of energy supply to the population is an example of it.

The AdaBoost algorithm creates an ensemble of weak classifiers that are called in a cascade, each subsequent tree corrects the errors of the previous one, and the predictions are based on a set of trees. The *AdaBoost* algorithm has been successfully used in research data in the field of healthcare [20], the lending [21], and detecting fraud with bank cards [22, 23].

Support vector machine

A *Two-Class Support Vector Machine* creates a binary classification model using a support vector machine algorithm. The Two-Class Support Vector Machine is a supervised learning algorithm for tagged data.

Neural networks

Classification using neural networks is a supervised learning method and therefore requires a tagged dataset that includes a tag column. A *Two-Class Neural Network* algorithm is used to predict binary outcomes, such as whether a patient has a specific disease, whether a machine might fail within a certain period of time, or whether there are deviations from a specific financial monitoring entity.

Bayesian networks

Bayesian networks are mainly used for solving diagnostic problems. For example, they are often used in medicine, credit scoring [24, 25], and in other tasks requiring risk assessment.

The *Two-Class Bayes Point Machine* algorithm uses a Bayesian approach to linear classification, it effectively approximates the theoretically optimal Bayesian mean for linear classifiers (in terms of generalization efficiency) by choosing one "mean" classifier, the Bayesian point.

Combined methods

Combining classification algorithms in solving practical problems can also show good results, with examples presented in studies [26–28].

The paper [29] proposes the concept of combining boost algorithms and support vector machines, which has shown high performance. In [30], the use of ensemble learning based on *XGBoost* for processing data from wearable sensors about the actions of the elderly was investigated.

The article [31] proposes a combined classifier using a support vector machine, PCA-based, and *k*-nearest neighbors algorithm methods for data processing in a gas analyzer. The work [32] studied the efficiency of ensembles of classifiers in the problems of predicting customer churn.

Anomalies detection

Another promising direction for solving financial monitoring problems is detecting anomalies. This group of methods has become widespread in modern conditions.

The problems of anomalies detection require knowing the set of values of the features $x^{(1)}, x^{(2)},\dots, x^{(m)}$, in order to find objects that are very different from others. The most popular anomaly detection tasks are identifying fraudulent transactions [22, 33–35], technical problems [36], detecting network intrusions, checking the values entered into the system, finding tax violators, etc.

A *One-class Support Vector Machine* creates a model that trains on data that has only one class, which is "non-anomalous". It displays the properties of "non-anomalous" cases and from these properties can predict which cases differ from the usual ones.

A *PCA-Based Anomaly Detection* algorithm analyzes the available data to determine what constitutes a "normal" class and applies distance metrics to detect anomalies [37, 38]. To detect anomalies, the projection of the values onto the eigenvectors is computed together with the normalized recovery error. Normalized error is used as an indicator of anomaly. The higher the error, the more anomalous the case is.

Classification algorithms are widely used for data analysis in medicine and healthcare [39–41].

Thus, we can conclude that any subject area has specific features, and the solution of practical problems requires the choice of a mathematical apparatus, and therefore this study is relevant.

RESEARCH METHODOLOGY

The sample of data for analysis included 334 credit institutions, 51 of them had their licenses revoked. The indicators of the bank reporting form No. 101 are investigated, in particular, the following groups of indicators are considered:

- investments in securities;
- investments in the capital of other organizations;
- loans to individuals;
- loans to enterprises and organizations;
- overdue debt;
- fixed assets and intangible assets;
- other assets;
- deposits of individuals;
- funds of enterprises and organizations.

The data slice is analyzed three months before the license revocation — based on the long-term practice of combating money laundering, this period was considered optimal, since, on the one hand, the deviant component of the bank's activities has time to manifest itself. clearly (it has been experimentally established that it *begins* to manifest itself about 6 months before the license is revoked), and on the other hand, there is still a sufficient margin of time for taking measures.

A number of experiments were carried out on data processing using binary classification algorithms:

- Two-Class Boosted Decision Tree;
- Two-Class Support Vector Machine;
- Two-Class Logistic Regression;
- Two-Class Decision Tree;
- Two-Class Neural Network;
- Two-Class Bayes Point Machine.

And for anomaly detection the One-Class Support Vector Machine and PCA-Based Anomaly Detection algorithms were used.

To determine the quality of the resulting models, we will use the *ROC* curve (Receiver operating feature) — this is a graph that allows us to assess the quality of a binary classification. The *ROC* curve displays the ratio between the proportion of objects from the total number of feature carriers, correctly classified as bearing a feature, and the share of objects from their total number, which do

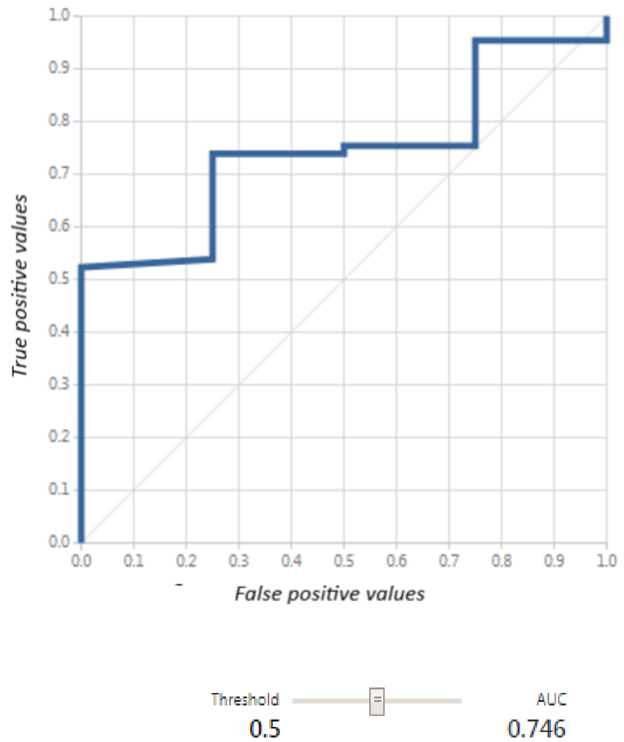


Fig. 1. **ROC curve and accuracy metrics for a Two-Class Boosted Decision Tree model**

Source: compiled by the author.

not carry a feature, erroneously classified as carrying a feature.

A quantitative interpretation of the *ROC* curve is provided by the *AUC* (area under the *ROC* curve) indicator — the area bounded by the *ROC* curve and the axis of the proportion of false-positive classifications [42–44]. The higher the *AUC*, the better the algorithm, the equality of the indicator 0.5 is equivalent to random fortune-telling.

The experiments were carried out on the *Microsoft Azure Machine Learning Studio* platform. *MS Azure ML* implements the capabilities of data analysis by various methods, including methods of classification, regression, cluster analysis, and anomaly search.

RESEARCH RESULTS

Let us demonstrate the classification of banks by prepared data using the two-class strengthened decision tree algorithm.

Fig. 1 shows the *ROC* curve of the constructed model; the value of the accuracy index AUC is 0.746.

Let's construct a Two-Class Support Vector Machine model for comparison (*Fig. 2*).

The *AUC* indicator for the constructed model is 0.619.

We will now build a Two-Class Logistic Regression and Two-Class Decision Tree models. The *ROC* curves are shown in *Fig. 3*.

The accuracy index of the *AUC* models is 0.588 for the Two-Class Logistic Regression model and 0.538 for the Two-Class Decision Tree, respectively.

Let's build a model of a Two-Class Neural Network. The *ROC* curve and accuracy index are shown in *Fig. 4*. The accuracy index of the *AUC* models is 0.654. We note that the neural network is built for three hidden layers. An increase in the number of hidden layers led to a deterioration in the quality indicators of the model's accuracy.

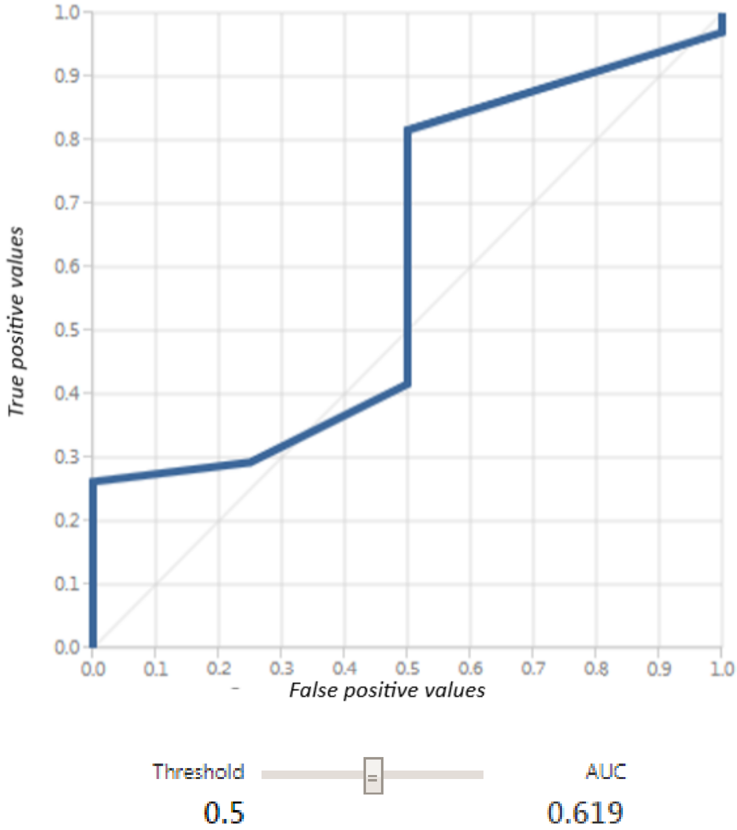


Fig. 2. **ROC curve and accuracy metrics for a Two-Class Support Vector Machine model**
Source: compiled by the author.

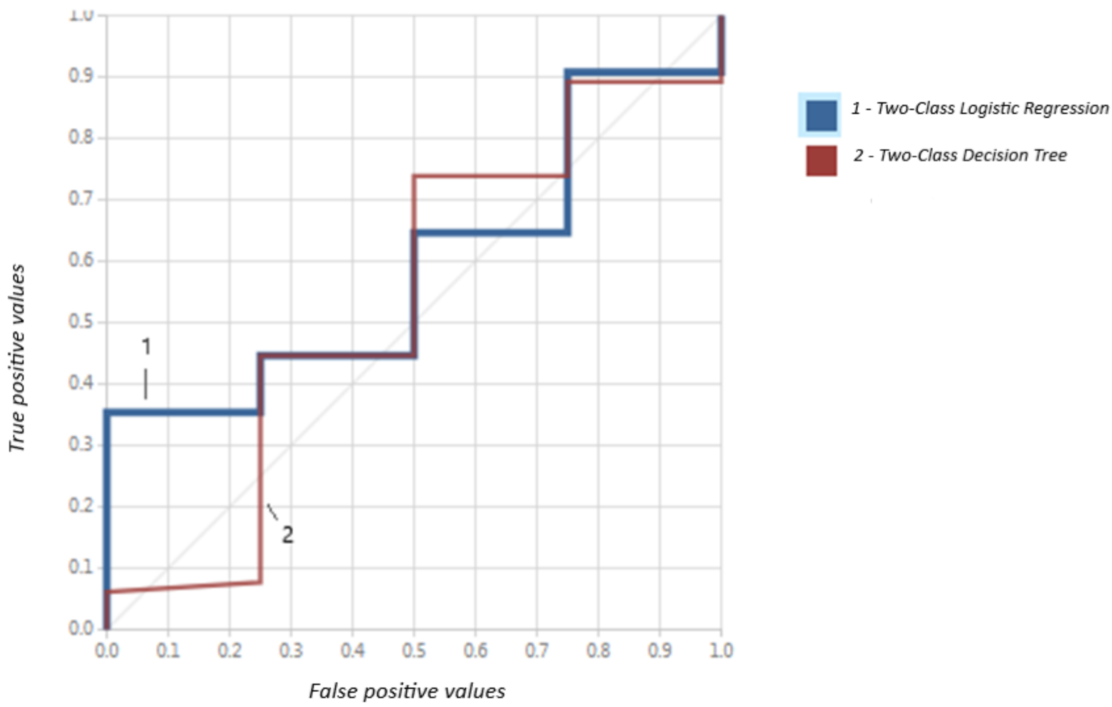


Fig. **ROC curves for Two-class Logistic Regression and Two-Class Decision Tree models**
Source: compiled by the author.

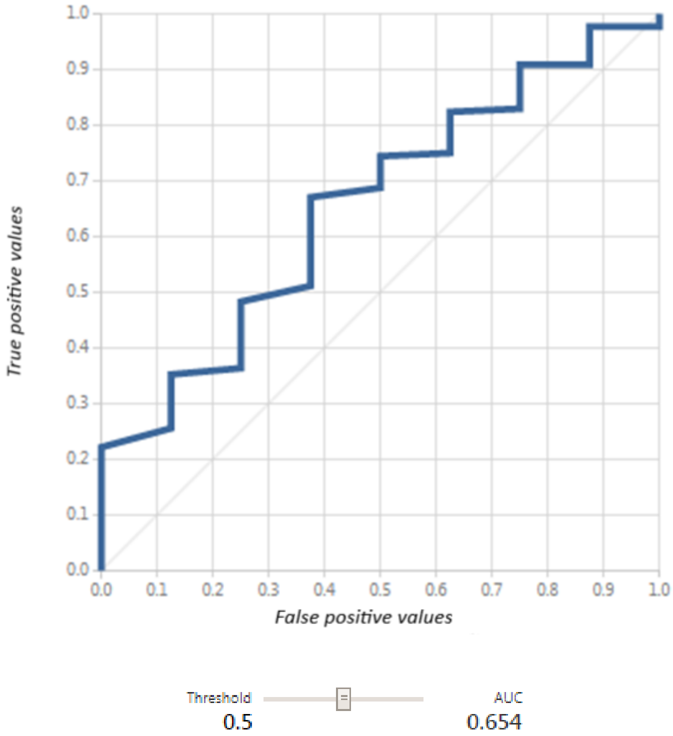


Fig. 4. **ROC curve and accuracy metrics for a Two-Class Neural Network**

Source: compiled by the author.

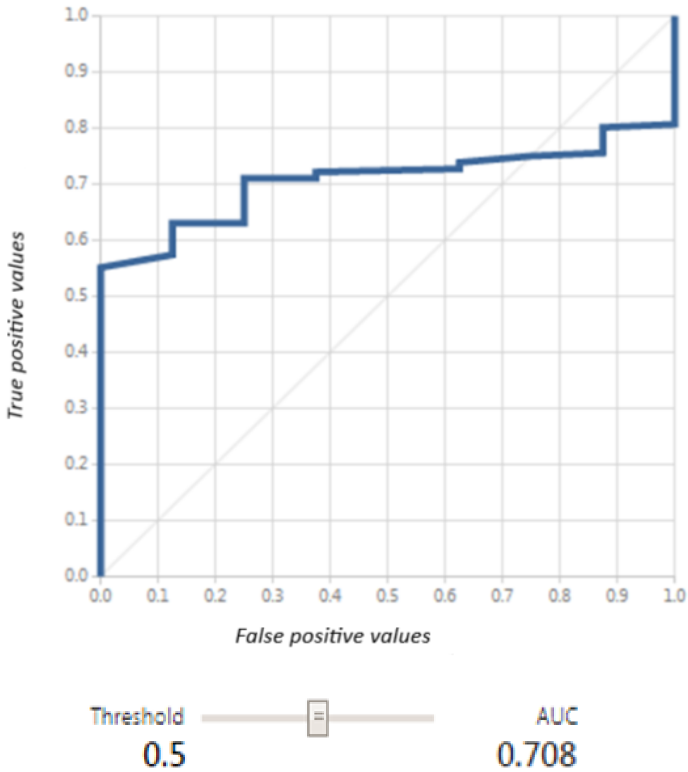


Fig. 5. **ROC curve and accuracy metrics for a Two-Class Bayes Point Machine**

Source: compiled by the author.

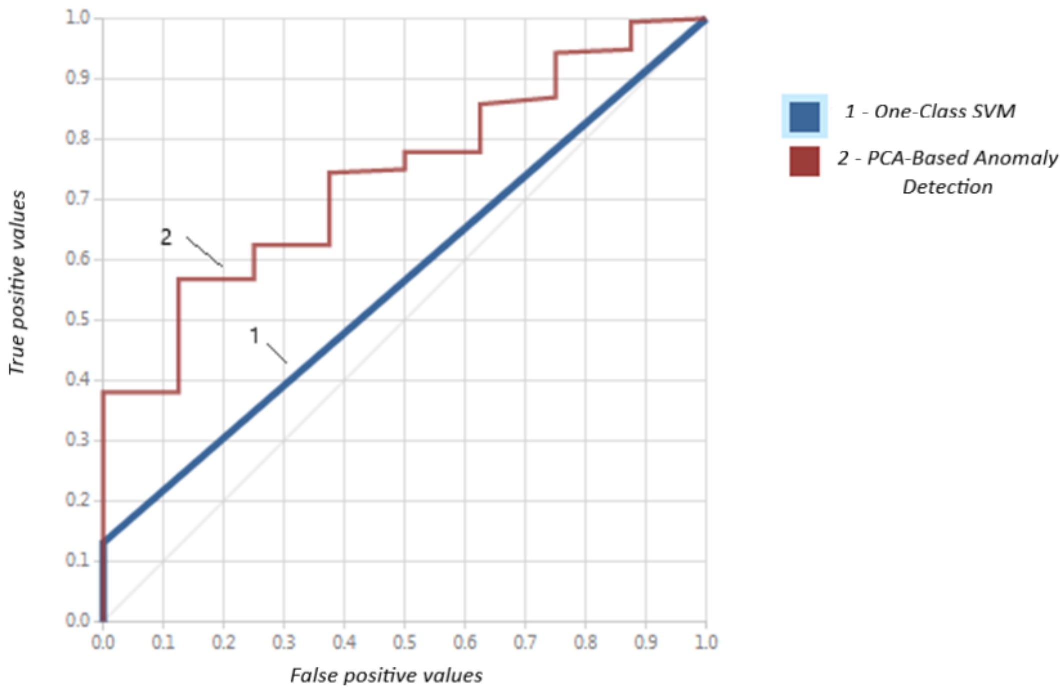


Fig. 6. **ROC curves for One-Class Support Vector Machine models and PCA-Based Anomaly Detection method**

Source: compiled by the author.

Let's build a model of a Two-Class Bayes Point Machine. The *ROC* curve and accuracy index are shown in *Fig. 5*. The accuracy index of the *AUC* models is 0.708.

Now let's look at anomaly detection algorithms — a One-Class Support Vector Machine and a PCA-Based Anomaly Detection algorithm. *ROC* curves for the One-Class Support Vector Machine and the PCA-Based Anomaly Detection models are presented in *Fig. 6*. The accuracy index of the *AUC* models is 0.683 for the One-Class Support Vector Machine model and 0.739 for the PCA-Based Anomaly Detection algorithm, respectively.

Table 1 shows the quality indicators of all models.

It can be concluded that the most accurate results were given by the Two-Class Boosted Decision Tree algorithm, and the PCA-Based Anomaly Detection algorithm method gives more accurate results compared to the One-Class Support Vector Machine algorithm.

Below are the predicted probabilities of license revocation obtained using the *Adaboost* algorithm. *Table 2* presents comparison results.

To assess the quality of the results, 50 credit institutions without deviations with a good reputation and 100 credit institutions knowingly involved in suspicious transactions were expertly selected. The greater number of objects with the lowest rating values is justified by the fact that in the tasks of financial monitoring, as a rule, it is required to find exactly deviant objects.

Of the 100 credit institutions identified as "deviant" (51 credit institutions had their licenses revoked within three months, another 46 banks had their licenses revoked within 4–10 months), *Adaboost* also classified 82 credit institutions as deviant.

Of the 50 lending institutions identified as "non-deviant", *Adaboost* classified 50 banks as "non-deviant".

Thus, the false positive rate for *Adaboost* will be:

False Positive Rate = $FP/(TN + FP) = 0/(50 + 0) = 0$.

The false negative rate will be:

False Negative Rate = $FN/(TP + FN) = 18/(82 + 18) = 0.18$.

Table 1

AUC metrics for algorithms for detecting deviations in the activities of credit institutions

	Anomaly detection algorithms		Classification algorithms					
	One-Class Support Vector Machine	PCA-Based Anomaly Detection	Two-Class Logistic Regression	Two-Class Decision Tree	Two-Class Boosted Decision Tree	Two-Class Neural Network	Two-Class Support Vector Machine	Two-Class Bayes Point Machine
AUC value	0.683	0.739	0.588	0.538	0.746	0.654	0.619	0.708

Source: compiled by the author.

Table 2

The results of the assessment of credit institutions (fragment of the table)

Bank	Has the license been revoked by the Bank of Russia	Is the bank deviant according to the AdaBoost algorithm
Sberbank	Not revoked	Non-deviant
VTB Bank	Not revoked	Non-deviant
Raiffeisenbank	Not revoked	Non-deviant
Bank GPB	Not revoked	Non-deviant
Alfa-Bank	Not revoked	Non-deviant
ROSBANK	Not revoked	Non-deviant
Rosselkhozbank	Not revoked	Non-deviant
…		
AO KB BRT	Revoked	Non-deviant
Bank RKB	Revoked	Deviant
AKB GAZSTROYBANK	Revoked	Deviant
KB TETRAPOLIS	Revoked	Deviant
KB MVKB	Revoked	Deviant
KB MEZHTRASTBANK	Revoked	Deviant
KB Prisko Capital Bank	Revoked	Deviant

Source: compiled by the author.

CONCLUSIONS

The article defines the specificity of the analysis of objects of financial monitoring, in particular, credit institutions, which is due to the high latency of deviant entities and their activities, the large volume and heterogeneous nature of information requiring analysis and interpretation, high requirements for professional knowledge and practical experience of expert analysts, and also significant time expenditures for the analysis of each individual subject.

Automation of the process of identifying unscrupulous credit institutions based on machine learning methods will allow regulatory authorities to promptly identify and suppress illegal activities.

In the field of financial monitoring, in order to make effective management decisions, it is necessary to promptly obtain objective assessments of economic entities (in particular, credit institutions), which, taking into account the specifics of the subject area, excludes a sequential manual check of the activities of entities. Automation of the process of identifying unscrupulous credit institutions based on machine learning methods will allow regulatory authorities to promptly identify and suppress illegal activities.

In order to automate the analysis of data on credit institutions, a mathematical toolkit has been selected that makes it possible to diagnose the involvement of a bank in the processes of money laundering. For this, a comparative analysis of the results of processing data on the activities of credit institutions was carried out using classification methods — logistic regression, decision trees (Two-Class Decision Forest algorithms, *Adaboost*), support vector machine (Two-Class Support Vector Machine algorithm), neural network methods (Two-Class Neural Network algorithm), Bayesian networks (Two-Class Bayes Point Machine algorithm) and anomaly detection — One-Class Support Vector Machine algorithm and PCA-Based Anomaly Detection algorithm.

The PCA-Based Anomaly Detection algorithm showed more accurate results compared to the One-Class Support Vector Machine algorithm. Of the considered classification algorithms, the most accurate results were shown by the Two-Class Boosted Decision Tree algorithm (*Adaboost*).

This confirms the hypothesis of the study on the possibility of diagnosing the involvement of a bank in the processes of money laundering using machine learning methods.

The above research results can be used by the Bank of Russia and Rosfinmonitoring to automate the identification of unscrupulous credit institutions.

REFERENCES

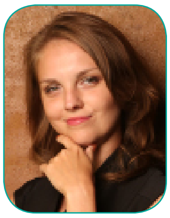
1. Kurkina E.P., Shuvalova D.G. Risk assessment: Expert method. *Problemy nauki.* 2017;(1):63–69. (In Russ.).
2. Zakharyan A.G. Expert assessment of the complex sustainability of a commercial bank. *Finansovye issledovaniya*. 2004;(9):14–19. (In Russ.).
3. Beketnova Yu.M., Krylov G.O., Denisenko A.S. The Problems of management and decision support in the government authorities on the example of the Rosfinmonitoring. *Informatizatsiya i svyaz' = Informatization and Communication*. 2018;(2):82–88. (In Russ.).
4. Klochko A.N., Logvinenko N.I., Kobzeva T.A., Kiselyova E.I. Legalizing proceeds from crime through the banking system. *Kriminologicheskii zhurnal Baikal'skogo gosudarstvennogo universiteta ekonomiki i prava = Criminology Journal of Baikal National University of Economics and Law*. 2016;10(1):194–204. (In Russ.). DOI: 10.17150/1996–7756.2016.10(1).194–204

5. Kononova N.P., Patlasov O. Yu., Kononov E.D. The risk-focused approach in the sphere of counteraction to laundering of the income and to financing terrorism. *Nauka o cheloveke: gumanitarnye issledovaniya = The Science of Person: Humanitarian Researches.* 2016;(2):183–189. (In Russ.). DOI: 10.17238/issn1998–5320.2016.24.183
6. Kuznetsova E.I., Burykin D.V., Masterova S.A. Risk-oriented internal control of credit institutions in the field of combating the legalization of proceeds from crime. *Vestnik ekonomicheskoi bezopasnosti = Vestnik of Economic Security*. 2017;(2):299–302. (In Russ.).
7. Pryakhin G.N., Ameleshin K.A. Improvement of methods of countering the legalization of criminal income and financing terrorism in the banking system. *Vestnik Chelyabinskogo gosudarstvennogo universiteta = CSU Bulletin.* 2019;(3):28–34. (In Russ.). DOI: 10.24411/1994–2796–2019–10304
8. Filatova I.V. Application of a risk-based approach to counteract the legalization (laundering) of proceeds from crime. *Vestnik Moskovskogo universiteta MVD Rossii = Vestnik of Moscow University of the Ministry of Internal Affairs of Russia.* 2019;(1):233–236. (In Russ.). DOI: 10.24411/2073–0454–2019–10055
9. Pavlidis N.G., Tasoulis D.K., Adams N.M., Hand D.J. Adaptive consumer credit classification. *Journal of the Operational Research Society*. 2012;63(12):1645–1654. DOI: 10.1057/jors.2012.15
10. Yap B.W., Ong S.H., Husain N.H.M. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*. 2011;38(10):13274–13283. DOI: 10.1016/j.eswa.2011.04.147
11. Khemais Z., Nesrine D., Mohamed M. Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *International Journal of Economics and Finance*. 2016;8(4):39. DOI: 10.5539/ijef.v8n4p39
12. Li Z., Tian Y., Li K., Zhou F., Yang W. Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*. 2017;74:105–114. DOI: 10.1016/j.eswa.2017.01.011
13. Louzada F., Anacleto-Junior O., Candolo C., Mazucheli J. Poly-bagging predictors for classification modelling for credit scoring. *Expert Systems with Applications*. 2011;38(10):12717–12720. DOI: 10.1016/j.eswa.2011.04.059
14. Siers M.J., Islam M.Z. Class imbalance and cost-sensitive decision trees: A unified survey based on a core similarity. *ACM Transactions on Knowledge Discovery from Data*. 2021;15(1):4. DOI: 10.1145/3415156
15. Bunkhumpornpat C., Sinapiromsaran K. Density-based majority under-sampling technique. *Knowledge and Information Systems*. 2017;50(3):827–850. DOI: 10.1007/s10115–016–0957–5
16. Devi D., Biswas S., Purkayastha B. A cost-sensitive weighted random forest technique for credit card fraud detection. In: The 10th Int. conf. on computing, communication and networking technologies (ICCCNT). (Kanpur, July 6–8, 2019). New York: IEEE; 2019. DOI: 10.1109/ICCCNT45670.2019.8944885
17. Zhang S. Multiple-scale cost sensitive decision tree learning. *World Wide Web.* 2018;21(6):1787–1800. DOI: 10.1007/s11280–018–0619–5
18. Zhu B., Baesens B., vanden Broucke S.K.L.M. An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*. 2017;408:84–99. DOI: 10.1016/j.ins.2017.04.015
19. Hurst W., Montanez C.A.C., Shone N. Towards an approach for fuel poverty detection from gas smart meter data using decision tree learning. In: Proc. 2020 3rd Int. conf. on information management and management science (IMMS). (London, Aug. 7–9, 2020). New York: Association for Computing Machinery; 2020:23–28. DOI: 10.1145/3416028.3416034
20. Al-Yarimi F.A.M., Munassar N.M.A., Al-Wesabi F.N. Electrocardiogram stream level correlated patterns as features to classify heartbeats for arrhythmia prediction. *Data Technologies and Applications*. 2020;54(5):685–701. DOI: 10.1108/DTA-03–2020–0076
21. Malhotra D.K., Malhotra K., Malhotra R. Evaluating consumer loans using machine learning techniques. In: Lawrence, K.D., Pai D.R., eds. Applications of Management Science. Bingley: Emerald Publishing Ltd; 2020;20:59–69. DOI: 10.1108/S0276–897620200000020004

22. Yang Y., Liu C., Liu N. Credit card fraud detection based on CSat-related AdaBoost. In: Proc. 2019 8th Int. conf. on computing and pattern recognition (ICCPR'19). (Beijing, Oct. 23–25, 2019). New York: Association for Computing Machinery; 2019:420–425. DOI: 10.1145/3373509.3373548
23. Tran P.H., Tran K.P., Huong T.T., Heuchenne C., Tran P.H., Le T.M.H. Real time data-driven approaches for credit card fraud detection. In: Proc. 2018 Int. conf. on e-business and applications (ICEBA 2018). (Da Nang, Feb. 23–25, 2018). New York: Association for Computing Machinery; 2018:6–9. DOI: 10.1145/3194188.3194196
24. Wu W.-W. Improving classification accuracy and causal knowledge for better credit decisions. *International Journal of Neural Systems*. 2011;21(4):297–309. DOI: 10.1142/S0129065711002845
25. Zhu H., Beling P.A., Overstreet G.A. A Bayesian framework for the combination of classifier outputs. *Journal of the Operational Research Society*. 2002;53(7):719–727. DOI: 10.1057/palgrave.jors.2601262
26. Marqués A.I., García V., Sánchez J.S. Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*. 2012;39(12):10916–10922. DOI: 10.1016/j.eswa.2012.03.033
27. Vukovic S., Delibasic B., Uzelac A., Suknovic M. A case-based reasoning model that uses preference theory functions for credit scoring. *Expert Systems with Applications*. 2012;39(9):8389–8395. DOI: 10.1016/j.eswa.2012.01.181
28. Akkoç S. An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*. 2012;222(1):168–178. DOI: 10.1016/j.ejor.2012.04.009
29. Turchetti Maia T., Pádua Braga A., de Carvalho A.F. Hybrid classification algorithms based on boosting and support vector machines. *Kybernetes*. 2008;37(9/10):1469–1491. DOI: 10.1108/03684920810907814
30. Wu Y., Qi S., Hu F., Ma S., Mao W., Li W. Recognizing activities of the elderly using wearable sensors: A comparison of ensemble algorithms based on boosting. *Sensor Review*. 2019;39(6):743–751. DOI: 10.1108/SR-11–2018–0309
31. Faleh R., Gomri S., Othman M., Aguir K., Kachouri A. Enhancing WO_3 gas sensor selectivity using a set of pollutant detection classifiers. *Sensor Review*. 2018;38(1):65–73. DOI:10.1108/SR-12–2016–0273
32. Lee S.-C., Faloutsos C., Chae D.-K., Kim S.-W. On detecting frauds in comparison-shopping services. In: Proc. 26th Int. conf. on world wide web companion (WWW'17 Companion). (Perth, Apr. 3–7, 2017). Geneva: IWWWC Steering Committee; 2017:811–812. DOI: 10.1145/3041021.3054219
33. Sohony I., Pratap R., Nambiar U. Ensemble learning for credit card fraud detection. In: Proc. ACM India joint int. conf. on data science and management of data (CoDS-COMAD'18). (Goa, Jan. 11–13, 2018). New York: Association for Computing Machinery; 2018:289–294. DOI: 10.1145/3152494.3156815
34. Lucas Y., Portier P.-E., Laporte L., Calabretto S., Caelen O., He-Guelton L., Granitzer M. Multiple perspectives HMM-based feature engineering for credit card fraud detection. In: Proc. 34th ACM/SIGAPP Symp. on applied computing (SAC'19). (Limassol, Apr. 8–12, 2019). New York: Association for Computing Machinery; 2019:1359–1361. DOI: 10.1145/3297280.3297586
35. Li Q., Xie Y. A behavior-cluster based imbalanced classification method for credit card fraud detection. In: Proc. 2019 2nd Int. conf. on data science and information technology (DSIT 2019). (Seoul, July 19–21, 2019). New York: Association for Computing Machinery; 2019:134–139. DOI: 10.1145/3352411.3352433
36. Ray S., Wright A. Detecting anomalies in alert firing within clinical decision support systems using Anomaly/Outlier Detection Techniques. In: Proc. 7th ACM Int. conf. on bioinformatics, computational biology, and health informatics (BCB'16). (Seattle, Oct. 2–5, 2016). New York: Association for Computing Machinery; 2016:185–190. DOI: 10.1145/2975167.2975186
37. Geiger B.C., Kubin G. Relative information loss in the PCA. In: Proc. IEEE information theory workshop (ITW). (Lausanne, Sept. 3–7, 2012). New York: IEEE; 2012:562–566. DOI: 10.1109/ITW.2012.6404738
38. Howard M.C. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*. 2016;32(1):51–62. DOI: 10.1080/10447318.2015.1087664

39. Khan H., Srivastav A., Mishra A.K. Use of classification algorithms in health care. In: Tanwar P., Jain V., Liu C.-M., Goyal V., eds. Big data analytics and intelligence: A perspective for health care. Bingley: Emerald Publishing Ltd; 2020:31–54. DOI: 10.1108/978–1–83909–099–820201007
40. Deepa B.G., Senthil S. Constructive effect of ranking optimal features using Random Forest, SupportVector Machine and Naïve Bayes for breast cancer diagnosis. In: Tanwar P., Jain V., Liu C.-M., Goyal V., eds. Big data analytics and intelligence: A perspective for health care. Bingley: Emerald Publishing Ltd; 2020:189–202. DOI: 10.1108/978–1–83909–099–820201014
41. Ramaswami G., Susnjak T., Mathrani A., Lim J., Garcia P. Using educational data mining techniques to increase the prediction accuracy of student academic performance. *Information and Learning Sciences*. 2019;120(7/8):451–467. DOI: 10.1108/ILS-03–2019–0017
42. Eskindarov M.A., Soloviev V.I., eds. Paradigms of the digital economy: Artificial intelligence technologies in finance and FinTech. Moscow: Cogito-Center; 2019. 325 p. (In Russ.).
43. Lee J.-S. AUC4.5: AUC-based c4.5 decision tree algorithm for imbalanced data classification. *IEEE Access*. 2019;7:106034–106042. DOI: 10.1109/ACCESS.2019.2931865
44. Zhang Q., Wang J., Lu A., Wang S., Ma J. An improved SMO algorithm for financial credit risk assessment — Evidence from China's banking. *Neurocomputing*. 2018;272:314–325. DOI: 10.1016/j.neucom.2017.07.002

ABOUT THE AUTHOR



Yuliya M. Beketnova — Cand. Sci. (Eng.), Assoc. Prof., Information Security Department, Financial University, Moscow, Russia
beketnova@mail.ru

The article was submitted on 25.12.2020; revised on 03.03.2021 and accepted for publication on 17.08.2021.
The author read and approved the final version of the manuscript.