Development of a methodology for detecting fraud with bank loans for legal entities

Nikita Tolstyakov, and *Natalia* Mamedova*

Plekhanov Russian University of Economics, 117997 Moscow, Russia

Abstract. This work is devoted to the development of highly efficient tools for making decisions by banking structures to issue and maintain a loan. The developed methodology is intended to support banking services for lending to legal entities. The paper presents the results of a comprehensive analysis of existing methods that can be used to identify scammers, the results of the analysis of data types for solving this problem and ranking them in terms of efficiency. A methodology for building algorithms for searching for scammers is proposed and the application of an algorithm for graph analysis of legal entity relationships for detecting fraud is demonstrated.

1 Introduction

The urgency of the task of improving the quality of decisions on the issuance of loans by banks to legal entities is easiest to prove with the help of numbers. For instance, in 2019, deliberate bankruptcy was detected in 5.5% of all completed bankruptcy cases, which, pursuant to experts, amounts to 181 billion rubles of creditors' claims [1]. In addition to deliberate bankruptcy, credit fraud and illegal obtaining of a loan cause significant damage to banks. The lower limit of damage for such cases in 2017 is estimated by experts at 1,076 RUR mln [2]. Traditionally, these offenses are recorded only after the bank has issued a loan.

The prospects for the return of all or part of the loan amount remain very vague, since they depend on a large number of factors. In 2019, the total amount of claims included in the register of creditors' claims amounted to 3301 RUR bln, of which only 146 billion rubles were satisfied. If we take for analysis only the creditors of the first stage, whose claims are satisfied in the first place, out of 809 RUR bln, only 321 RUR bln were paid, which is 39%. Herewith, out of 36354 bankruptcy cases, signs of intentional bankruptcy were identified in 2,023 cases (5.5%), in 5349 (14%) cases, there is not enough information about the conclusion [1]. Analysis of the above statistics shows that the main damage from these crimes is borne by the banking system.

In this paper, the term "fraud" means a combination of three types of crimes: credit fraud, illegal obtaining of a loan, as well as the deliberate bankruptcy. Their combination is due to the fact that these crimes are united by a cause-and-effect relationship and are usually classified in the aggregate. In addition, the proposed solutions are a universal tool for detecting any of these crimes.

* Corresponding author: nmamedova@bk.ru

You may feel that organizations such as banks are armed with highly efficient tools that protect them from the risk of lending to a fraudulent entity. However, as the analysis of the subject area shows, this is not the case.

For instance, it is quite obvious that the key points in time for the most efficient fixing of fraud are: the moment of making a decision to issue a loan and the moment of deciding on a strategy for settling the problem debts. It is also a common point of view thcurrently of making a decision to grant a loan is provided with tools in all modern banks that use the methodological recommendations of Basel II [3]. And, perhaps, many are familiar with Deloitte models based on international standards (IFRS 9) and practices for modeling expected credit losses - we mean such models as Probability of default (PD); Loss Given Default (LGD); Expected credit losses (ECL) 4, 5]. And literally on hearing modern trends in the field of artificial intelligence and machine learning to develop a solution to the problem of credit scoring [6-8]. However, with all this, the problem of finding and implementing highly efficient fraud prevention tools continues to be relevant.

To date, there are very few researches in which methods and algorithms are proposed for identifying scammers among legal entities with bank loans [9-11]. The world practice of using financial methods to detect fraud includes such common methods as:
- identification of an unsatisfactory balance sheet structure based on a system of criteria for assessing a possible bankruptcy;
- E. Altman's model "Z-score" [12] and other similar models based on stochastic factor analysis and deterministic factor analysis - models of J. Conan, J. Lego, R. Fox, G. Springgate, R. Tafler, G. Tishaw et al. [13-15];
- methods that are based on the Argenta indicator (A-account), proposed by George Argenti [16].

Thanks to the above models and methods, banks can reduce the risks of fraud or illegal obtaining of a loan at the time of making a decision to grant it. And usually the projection of such models and methods is a decrease in the volume of loans issued by the bank. And, if an unscrupulous borrower made a decision about deliberate bankruptcy or fraud after receiving a loan, banks in this case are even less protected. Such cases are investigated after the damage has been caused to the bank, and in most cases the damage cannot be compensated, as evidenced by the statistics given earlier.

The bank makes a decision on granting a loan on the basis of information provided to it by a legal entity. In this case, the legal entity itself can be fictitious. Not being able here to describe in detail the list of deliberately false and (or) inaccurate information provided to the bank, as well as signs of fictitiousness of a legal entity, we will only indicate that they are divided into two groups: information about the economic situation and information about the financial condition, as well as indicate the source where one can find detailed information.

Let us also admit a situation when, after a loan has been issued, a legal entity violates credit discipline. Herewith, fraud consists in persuading the bank on the basis of deliberately false and (or) inaccurate information to make a decision on the application of a credit strategy (debt restructuring), and not a default strategy (initiation of bankruptcy proceedings).

Based on the above analysis of the subject area, we will formulate the task of developing a methodology for detecting fraud with bank loans for legal entities. For the correct formulation of the problem, it is necessary to select the exact segment for which this infrastructure will be built. As part of this work, a methodology will be proposed for creating an IT infrastructure for a large bank that needs to use Big Data technologies to store and process information about borrowers.

It is also worth mentioning that this work assumes that general recommendations for reducing the risk of fraud are followed. The most important of these is monitoring employees for collusion with a scammer. If it is not executed, then the best system will not be able to prevent the loss of money, although it will help to identify the attacker.

Thus, our task was to develop a methodology for detecting fraud (hereinafter referred to as the Methodology) at the stage of analyzing a loan application, at the monitoring stage, if there is little internal data about the borrower, and at the monitoring stage, if the internal data on the borrower is sufficient to build a conclusion on their basis. For this, machine learning methods and other modern and efficient mathematical algorithms were used, which is justified by their efficiency in solving data-based problems. A method was developed for storing and transferring the analysis results directly to an employee for further work with them. The IT architecture of the system, created on the basis of the developed Methodology, should ensure the interaction of this system with external and internal data sources.

2 Study methodology

The development of the Methodology began with the selection of data sources for identifying scammers among legal entities. In this paper, it is proposed to build algorithms for searching for scammers, as much as possible based on the internal data of the bank, then based on data from third-party organizations, and last but not least, basing their conclusions on the data provided by the borrower himself. This approach will allow to create a system that is maximally protected from external interference.

As the internal data of the bank, the key data became the borrower's transactions and information about changes in his accounts. This approach will help to identify the counterparties of the borrower and rank them in order of importance. It will help to find hidden connections between various borrowers of the bank and other companies that are not formally related. The following data were used as data on the borrower from third-party sources: The database of the Federal Tax Service (including the Unified State Register of Legal Entities and the Unified State Register of Legal Entities), the National Bureau of Credit Histories, the Unified State Register of Real Estate, the Unified State Register of Bankruptcy Information (EFRSB), and arbitration case files.

Maximizing the security of data sources should be the first priority when building fraud detection systems. But one should not give up at all from the analysis of "unreliable" data. In them, one can identify patterns and use them in the future. The bank usually has a large number of independent data sources. A fairly large part of these sources is open source, which dramatically reduces the cost of developing decision support systems. Assessing the quality of a source should consider the reliability and validity of the data in it, the option of ranking sources by a set of criteria. It is also worth noting that if the bank for some reason does not store the history of internal data about the user, then he needs to start doing this, since they are the most valuable and reliable source of information.

The next stage in the development of the Methodology provided for the organization of the process of building algorithms for finding scammers among legal entities. Three different algorithms were built:

1)Algorithm for detecting fraud based on the purpose of the payment in the transaction.

2)Fraud detection algorithm based on RAS reporting.

3)Algorithm for graph analysis of the borrower's connection to detect fraud.

Since the development of the algorithms was carried out by a machine experiment, the purpose of which is to establish the quality of the algorithm, we will designate the main parameters used to create the dataset:

1) For each experiment, a set of data was collected from: recognized scammers, borrowers with a high credit rating, borrowers without signs of fraud, for which there was a delay (more than 90 days) or bankruptcy proceedings were initiated.

2) The data was collected for the time period from 06/01/2014 to 01/01/2020.

3) When creating a sample of scammers, all three types of fraud were considered in equal proportions.

4) Companies belonging to the segments of micro, small business and medium, large business are taken in the proportion of 80% to 20%, respectively.

5) Borrower fraud data is collected as of the earliest known fraud period.

6) If it was impossible to calculate the algorithm pursuant to the borrower's data collected for the selected period, then this borrower was ignored.

7) For a borrower with a good credit rating, data are selected at random from the time period.

8) For borrowers who defaulted or for whom bankruptcy proceedings began, data was collected at the time of delinquency or bankruptcy.

Not being able to present here all the constructed methods, due to the limitation of the volume of the article, we limited ourselves to the presentation of one of the algorithms - the Algorithm of graph analysis of the borrower's connection to detect fraud. The choice in favor of presenting this particular algorithm in this work is due to the fact that it uses data from third-party organizations. In addition, this algorithm is interpretable, which allows the employee, who will conduct the analysis, not only to accept the fact that the criterion of fraud has been triggered for the borrower, but also to study the problem in more detail and in detail, which will also reduce the likelihood of false positives. However, it should be noted that the use of a combination of the developed algorithms will give the greatest cumulative effect.

The algorithm for graph analysis of the borrower's connection to detect fraud is designed to use third-party data in the first place. The algorithm takes into account the approach to finding scammers on various online trading platforms and uses Markov networks to classify the borrower. The general idea of this algorithm is to classify the borrower based on his relationships with other companies. The idea behind this algorithm is based on several identified patterns:

1) Most often, in case of fraud, a borrower tries to withdraw credit money through related counterparties who are accomplices.

2) A scammer can enhance his financial performance with the help of related parties. There may be fictitious transactions in which one company transfers money to another, but does not receive anything for it.

3) Selling property at a lower price in order to commit deliberate bankruptcy is more likely to happen through related companies.

Herewith, the connection between the scammer's company and the company that helps him can be based on: close, family ties between the owners of the companies; the owner of the fraudulent company may be the beneficiary of the counterparty; they can belong to the same group of companies; a fraudulent company may own part of a counterparty's company. Less commonly, companies that are weakly connected, but act as ordinary accomplices, enter into conspiracy.

For a correct understanding of the operation of the algorithm, it is worth clarifying that all types of fraud have been transferred into one general class "scammer". As a consequence, the problem of not multiclass classification will be solved, but the problem of binary classification. This approach is based on two factors. The first factor - the proximity of fraud methods in cases of deliberate bankruptcy, credit fraud and illegal obtaining of a loan makes these offenses difficult to separate without in-depth analysis of the causes and consequences of the offense. The second factor is the fact of an unbalanced sample. In such cases, it is common to use a one-against-all approach. With this approach, all but one of the sample classes are combined into one, and the problem is reduced to a binary classification. After the new classification problem has been solved, the grouped class is split back, and again it is necessary to solve the multiclass classification problem, but the sample becomes one class less. Another argument in favor of the approach is the following: if any of the above types of fraud are detected, the most efficient solution for the bank will be either not to issue a loan or start bankruptcy proceedings.

3 Research results

It is necessary to build an algorithm that can build connections of various types between companies. The classical way to solve such problems is to use graph-based algorithms. The algorithm presented below is based on the ideas of analyzing fraudulent accounts on social networks and on electronic trading platforms.

Algorithms for identifying fraudulent accounts assume the division of all accounts into three categories: scammers, accomplices, and ordinary users. Therefore, it is necessary to break down all borrowers and their associated persons into these three categories. Obviously, it is necessary to add directly to the category of scammers the companies that commit the offense. Then those companies with the help of which the fraud is committed will be considered accomplices. All other companies will be transferred to the class of ordinary borrowers or the class of "unknown" if there is too little information on them.

The next step that should be taken is to select directly the types of connection, along which the graph will be built. Based on the previously presented patterns, it is proposed to distinguish the following types of connections, sorted by their strength and each such connection, we assign a certain weight w from 0 to 1:

1) Identified fraudulent transactions found “manually” by bank employees. The weight w is 1.

2) One company owns part of another company. The weight w is 0.8.

3) The beneficiary of one company is the beneficiary of another company or a relative of the beneficiary. The weight w is 0.6.

4) The legal, postal or office addresses of the companies are the same. The weight w is 0.3.

5) There is information that money transfers are being made between companies. The weight w is 0.05.

After choosing the types of communication on their basis, it is necessary to build an undirected graph, on which, first, the companies may be marked by classes, which will increase the quality of prediction. It should be noted that the markup should not be based on assumptions, but should be based on factual information about the category of the borrower.

The algorithm itself is based on the Markov network [17, 18]. A Markov network is a graph model in which a set of random variables has the Markov property. In more detail, the essence of the algorithm is to determine the category of the graph vertex (company category) based on information about the category of its neighbors and its own current state. To categorize a company, it is necessary to update the information about the state of the vertices using the Belief Propagation method [19], which is presented in the formulas:

$$mij(\sigma) \leftarrow \sum\sigma F\ \psi(\sigma\grave{O},\ \sigma) * \sum l\in L\ wl * \prod n\in N(i)/j\ mnj(\sigma\grave{O}) \quad (1)$$

$$bi(\sigma) \leftarrow k \prod j\in N(i)\ mij(\sigma) \quad (2)$$

Here $mij(\sigma)$- message sent by node i to node j; N(i) – set of nodes adjacent to i; $\psi(\sigma\grave{O},\ \sigma)$ – entry into the propagation matrix giving the probability of being in the state $\sigma\grave{O}$ in the presence of a neighboring node in the state σ; wl – weight l of the type of connection between vertices; L – set of all connections between two vertices; k – normalization constant; $bi(\sigma)$ – the level of trust in node i to the state σ.

The Belief Propagation matrix is a matrix that describes the probabilities of transition from one state to another based on the states of neighboring vertices, it is shown in Table 1. Now, based on the level of confidence in the node $bi(\sigma)$, it is necessary to assign it one of the categories, for this on the basis of maximizing the classification quality function F -measures need to choose a threshold value r.

It should be noted that the values of the function $\psi(\sigma\grave{O},\ \sigma)$ and the weights w were expertly and require correction for each user of the algorithm. These weights can be obtained

by constructing a graph pursuant to the given rules, but with reliably known categories of companies and calculating the probability based on Bayesian statistics or on the basis of an optimization problem.

Table 1. Trust extension matrix.

Adjacent connection	Node status		
	Swindler	**Accomplice**	**Fair**
Swindler	0.4	0.9	0.2
Accomplice	0.8	0.4	0.3
Fair	0.2	0.3	0.6

To determine the quality of the proposed algorithm and build the trust propagation matrix, a machine experiment was carried out. For the experiment, the Python language was chosen. In particular, the SciPy library, created for solving algebra and optimization problems, was used as the main tool. A sample was collected to check the quality of the algorithm, which included transactions of 1000 borrowers who were found to be overdue for more than 90 days, 1000 borrowers with high credit ratings and 198 borrowers who were scammers whose crime was proven in court.

Further, the sample was divided into training and test samples. The first step was to combine 500 overdue borrowers and 99 scammers into one sample. For each, all his transactions were split into 5 equal time intervals. The second step is to check the quality of the remaining 500 late borrowers and 99 scammers. As a final step, 1000 borrowers with good credit history were tested. Since in this part of the sample there are “non-crooks”, i.e. there is only one class. Therefore, for this quality assessment, the Accuracy metric was used, which shows the ratio of correctly predicted objects to all objects. A test with "non-cheaters" is necessary to fully assess the quality of the model. It will show how many good borrowers have been classified as fraudulent. If this number turns out to be unacceptable, then it is necessary to abandon this algorithm.

To find the optimal values of the trust propagation matrix, it was decided to use the optimization problem. At the first stage, N graphs were built from the borrowers of the training sample and all classes of companies were marked on these graphs, where N is the number of companies in the training sample. For the construction, all the previously indicated types of communication between borrowers were indicated. Since the number of connections of the borrower with the help of transactions can be extremely large, it was customary to limit the size of the graph. To do this, we cut off all the links of the borrower considered in this column that were more than 3 steps away from him.

Then N similar graphs were built, but the class of the borrower was not marked on them. After that, an optimization problem was built, in solving which such values of the trust propagation matrix should be selected, at which the error in predicting the borrower’s class on the pre-labeled graph and the graph that labels the algorithm was minimal. In optimization problems, it is customary to minimize the value of a certain quality function. In this case, to find the optimal values of the trust propagation matrix, we used the LogLoss [20] metric, which is the standard optimization metric for classification problems.

4 Discussion of results

For the final check of the algorithm quality on the test sample, Out-of-Time testing was carried out. To do this, good borrowers used 5 different bond states for different periods of the company’s existence. For the scammers, a similar split took place, but the interval in which the fraud was committed was used as the time period for the split.

To test the ability of the algorithm to predict fraud for a period in which there are not enough transactions to build a fraud search algorithm based on the purpose of the payment,

a similar test was carried out, but the link was removed for transactions that were recognized as fraudulent.

To test the ability of the algorithm to work without communication for ordinary transactions, one more testing was carried out, similar to the previous one, but in this case, the relationship for ordinary transactions was removed. Such testing is essential. If the results of this testing do not differ from the results with all links, then the link for regular transactions can be deleted. Which in turn will make the algorithm computationally simpler.

The following testing was carried out with the removal of all types of transactions. This is necessary to check the quality of the algorithm currently of making a decision to issue a loan, when there is no information about transactions. The test results are shown in Table 2.

As a result of the execution of this algorithm, knowledge was obtained about the category of the company at a given moment in time, during the construction of which, in contrast to the usual algorithms based on the Markov network, a very important knowledge was introduced about the strength of the type of connection between the two companies. Obviously, the presence of transactions between scammer A and some company B does not make company B an accomplice if there is only a transactional relationship between them, but the likelihood that company B is an accomplice increases sharply if some of the transactions were found to be fraudulent.

Table 2. Borrower classification quality based on graph analysis.

Classification signs	Precision	Recall	F-measure
All types of communication	0.61	0.64	0.69
Transactions recognized as fraudulent are excluded	0.52	0.49	0.6
Communication on regular transactions is excluded	0.57	0.51	0.64
All types of transactions are excluded	0.46	0.44	0.45

Table 2 shows the quality of this algorithm, and Figure 1 shows the visualization of the algorithm in two states: the initial state without entering a priori information about the company category (a) and the result of the algorithm (b), where swindlers are marked in red, accomplices in yellow, and white "good" companies, and gray companies for which the result is not known.

Thus, an algorithm was presented to search for scammers and those companies with the help of which the offense occurs. The presented algorithm is less efficient than an algorithm based on a neural network of the RNN-Autoencoder type, but is similar to an algorithm based on time series analysis. Herewith, the RAS reporting algorithm and this algorithm perform different tasks. The first tries to find scammers who falsify documentation, the second tries to find scammers who use other companies in their actions.

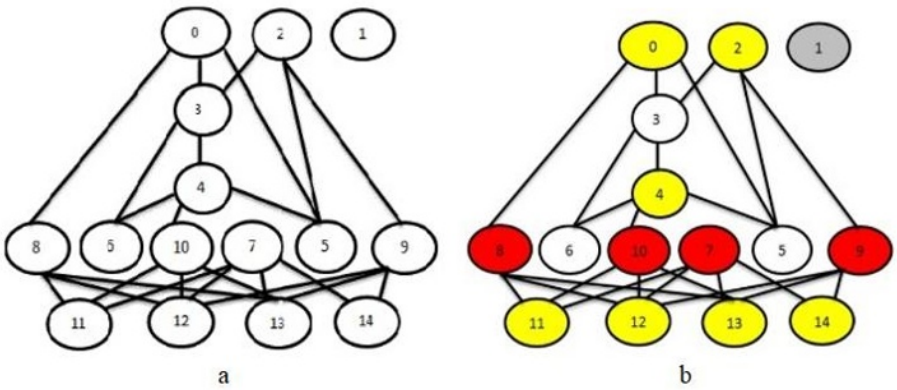

Fig. 1. Graph model before (a) and after applying algorithm (b) (creately.com).

The description of the graph analysis algorithm for the client's connection to detect fraud was given in sufficient detail for the implementation of software based on it. It is worth noting that this method has great flexibility in terms of choosing the main influencing factors, each user of this algorithm will be able to customize the types of links and enter the coefficients that seem correct to him. Due to this flexibility, this algorithm shows flexibility in the resources necessary for its use, because a company usually does not have very many legal ties, it can also be limited by the number of vertices that will be considered in the calculation.

Moreover, this algorithm is not limited in time, like the other two proposed algorithms. It can also be used for companies that have just opened, since most of the signs are based on information that has been in the public domain from the very beginning of the existence of a legal entity.

5 Conclusion

Summarizing the results of the research process and the results obtained, the following should be indicated.

There is a comprehensive description of the subject area. The state of the problem of fraud with bank loans among legal entities and the procedure for using machine learning algorithms to identify scammers in this area are presented. The statement of the problem includes the development of a methodology for detecting fraud with bank loans among legal entities. Three algorithms are proposed that demonstrate different approaches to solving the problem, and the procedure for building one of them, an algorithm for graph analysis of the borrower's connection for detecting fraud, is described.

The practical significance of the work lies in the fact that the results obtained provide the construction of an independent or integrated system to identify fraud with bank loans among legal entities. The scientific novelty of the research lies in the development of unique algorithms for detecting scammers. The proposed solutions in the complex are aimed at reducing the bank's losses from the risk of fraud in the provision of lending services to legal entities.

References

1. Fedresurs.ru, Rezul'taty protsedur v delakh o bankrotstve za 2019 god (2020). https://fedresurs.ru.
2. A.B. Ivanova, Vestnik ekonomiki, prava i sotsiologii **3**, 83 (2018)
3. S. Li, Emerging Trends in Smart Banking: Risk Management Under Basel II and III, IGI Global (2014)
4. M.D. Ermolova, G.I. Penikas, Model assisted statistics and applications **4**, 335 (2017)
5. J. Eckert, K. Jakob, M. Fischer, J. of Credit Risk **12(1)**, 97 (2016)
6. F. Louzada, A. Ara, G.B. Fernandes, Surveys in Operations Research and Management Science **21(2)**, 117 (2016)
7. R.Y. Goh, L.S.Lee, Advances in Operations Research **9**,1974794 (2019)
8. G. Teles, J.J.P.C. Rodrigues, K. Saleem, S. Kozlov, R.A.L. Rabêlo, Machine learning and decision support system on credit scoring Neural Computing and Applications, **32(14)**, 9809 (2020)
9. Yu.K. Ivanova, Beneficiar **46**, 22 (2019)
10. D. Rahmawati, R. Sarno, C. Fatichah, D. Sunaryono, *3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, 35 (2017)

11. D. Sarma, W. Alam, I. Saha, M.J. Alam, S. Hossain, *Proceedings of the 2nd International Conference on Inventive Research in Computing Applications*, ICIRCA 2020, 9182954, 642 (2020)
12. E. I. Altman, M. Iwanicz-Drozdowska, E.K. Laitinen, A. Suvas, Financial Management and Accounting **28(2)**, 131 (2017)
13. N. M. Chuong, P. G. Ciarlet, P. Lax, D. Mumford, D.H. Phong, *Advances in Deterministic and Stochastic Analysis* (2007)
14. N. Bărbuță-Mișu, V. Mazilescu, Risk Governance and Control: Financial Markets and Institutions **1(1)**, 112 (2011)
15. A. Agarwal, I. Patni, Int. J. of Innovative Technology and Exploring Engineering **8(6)**, 131 (2019)
16. A. Tikhomirov, E. Skripka, *Proceedings of the 33rd International Business Information Management Association Conference, IBIMA 2019: Education Excellence and Innovation Management through Vision 2020*, 1817 (2010)
17. S. Liu, K. Fukumizu, T. Suzuki, Behaviormetrika **44(1)**, 265 (2017)
18. M. Matalytski, D. Kopats, Probability in the Engineering and Informational Sciences **35(1)**, 158 (2021)
19. X. Lin, D. Niu, X. Zhao, B. Yang, C. Zhang, Neurocomputing **325**, 131 (2019)
20. H. Brink, J. W. Richards, M. Fetherolf, *Real-World Machine Learning, Manning Publications* (2016)