DOI: 10.17323/1998-0663.2019.3.52.66

# Comparative analysis of methods for forecasting bankruptcies of Russian construction companies

**Alexander M. Karminsky**
E-mail: karminsky@mail.ru

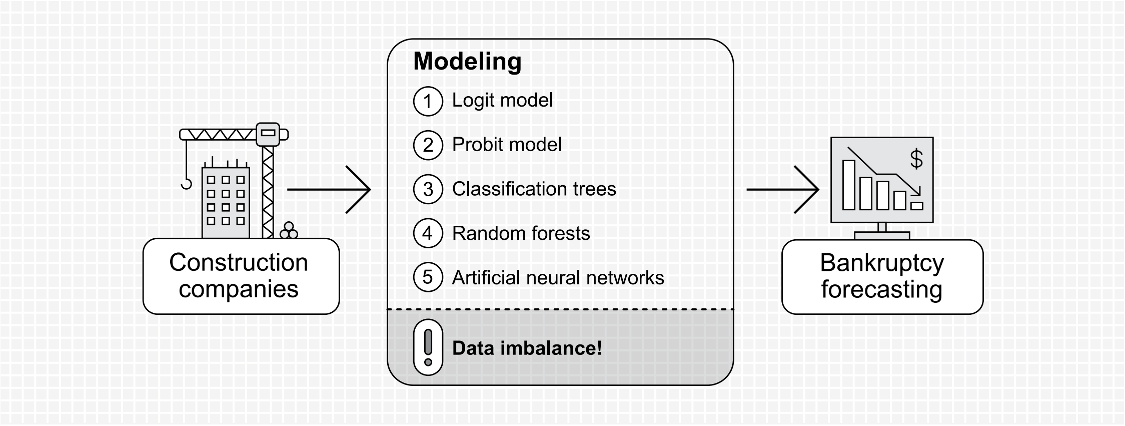
**Roman N. Burekhin**
E-mail: romanvia93@yandex.ru

National Research University Higher School of Economics
Address: 20, Myasnitskaya Street, Moscow 101000, Russia

**Abstract**

This paper is devoted to comparison of the capabilities of various methods to predict the bankruptcy of construction industry companies on a one-year horizon. The authors considered the following algorithms: logit and probit models, classification trees, random forests, artificial neural networks. Special attention was paid to the peculiarities of the training machine learning models, the impact of data imbalance on the predictive ability of models, analysis of ways to deal with these imbalances and analysis of the influence of non-financial factors on the predictive ability of models. In their study, the authors used non-financial and financial indicators calculated on the basis of public financial statements of the construction companies for the period from 2011 to 2017. The authors concluded that the models considered show acceptable quality for use in forecasting bankruptcy problems. The Gini or AUC coefficient (area under the ROC curve) was used as the quality markers of the model. It was revealed that neural networks outperform other methods in predictive power, while logistic regression models in combination with discretization follow them closely. It was found that the effective way to deal with the imbalance data depends on the type of model used. However, no significant impact on the imbalance in the training set predictive ability of the model was identified. The significant impact of non-financial indicators on the likelihood of bankruptcy was not confirmed.

**Graphical abstract**



**Key words:** bankruptcy; the construction sector; imbalance data; machine learning models; parametric models of prediction of bankruptcy.

**Citation:** Karminsky A.M., Burekhin R.N. (2019) Comparative analysis of methods for forecasting bankruptcies of Russian construction companies. *Business Informatics*, vol. 13, no 3, pp. 52–66. DOI: 10.17323/1998-0663.2019.3.52.66

## Introduction

In a market economy, forecasting financial insolvency is an important task for any company. To achieve this goal, different methods of assessing credit risks are used. Their purpose is to proactively and effectively forecast the onset of an adverse situation in the company. Typically, these methods are parametric models characterized by a relatively simple mathematical apparatus and a simple qualitative interpretation. These methods are static, do not take into account subtle economic or behavioral factors, and the predictive ability of the models decreases with the nonlinear nature of the relationships between the indicators.

Market models (structural models and shortened models) are often too complex or market dependent. To apply them, you need access to a large amount of data (market value of share capital, debt obligations, spreads of bond yields, etc.) Despite the widespread use of market models by Western companies, their use in the Russian market is difficult due to the small number of listed securities. To conduct an effective credit policy, new methods must be flexible and adaptable to the changing realities of a market economy. Therefore, there is currently an interest in models based on machine learning algorithms, including classification trees, random forests, gradient boosting, artificial neural networks, etc.

There are a number of common problems associated with predicting bankruptcy of companies. Firstly, the economic indicators describing the state of the company differ in various studies, and their integration into the most effective model causes additional difficulties. Secondly, there is a problem of data imbalance, since there are more solvent companies than bankrupt ones. As a result, the trained model tends to classify companies as reliable, although they may have signs of financial failure. Thirdly, the very concept of “bankruptcy” can be interpreted in different ways, so different companies can fall into this category. In this work, the category of bankrupts includes companies in respect of which legal bankruptcy proceedings have begun, as well as companies that have liquidated voluntarily.

Despite the importance of the task of forecasting bankruptcies using more advanced methods, there are not so many domestic works in this area, and works on forecasting reviews of bank licenses are more likely to be the exception [1, 2]. A feature of this work is the comparison of regression models and models based on machine learning methods in the tasks of predicting bankruptcies of companies based on one industry. Considerable attention is paid to the specifics of building machine learning models, the impact of data imbalances, as well as non-financial indicators on the predictive ability of models.

The construction industry is a link between other industries, which determines its importance in the national economy. Today in Russia there are more than two hundred and seventy thousand companies performing certain construction work (design, engineering calculations, construction, etc.). Their number, as well as the high level of defaults in this sector makes it difficult to choose a suitable partner. This industry is one of the most affected by the crisis. In particular, the volume of work in comparable prices has not ceased to fall since 2014, and by the end of 2017, construction turned out to be an industry with one of the highest share of bad debts. Lending to the construction sector represents a significant part of the Russian banking business. Therefore, an increase in the number of insolvent construction companies can cause instability in the banking sector. Moreover, national and international regulatory requirements (recommendations of the Basel Committee) force the use of an advanced approach based on internal ratings to quantify risks in order to reduce the burden on capital. Therefore, the problem of forecasting the future state of construction companies is relevant, and new tools for forecasting bankruptcies are in demand.

This paper answers the question whether models based on machine learning methods can be a worthy alternative to regression models when applied to the field of bankruptcy forecasting of companies in the non-financial sector, using the construction industry as an example. It is concluded that all the considered models are capable of predicting bankruptcy in the next 12 months, while neural networks are superior to other methods in identifying insolvent companies, and logistic regression models combined with discretization closely follow them. A negative effect of the imbalance of the training set on the predictive ability of the model was not found[1].

## 1. Models for predicting financial insolvency

Regression models (logit and probit models) are common in the problems of identifying solvent and insolvent borrowers [3]. Their advantage lies in the absence of severe restrictions on functioning, ease of interpretation and simplicity of calculations. An important drawback of these models is a decrease in prognostic ability with the non-linear nature of the relationships between the indicators, while machine learning algorithms are less sensitive to these problems. There are many works proving the possibility of using advanced methods for predicting company insolvency [4–7].

The authors [8] were among the first to use classification trees to predict company bankruptcies. They found that their classification trees outperform discriminant analysis. It was

[1] Preliminary results of the study were presented in the graduate work by Roman N. Burekhin, performed at the HSE Faculty of Economic Sciences in 2018

also noted that with the complication of the model (inclusion of a larger number of factors), its accuracy deteriorated due to overfitting. However, this success did not cause the widespread use of decision trees in this area. In the future, in most works, there is a comparison of the effectiveness of decision trees with other algorithms. The random forest algorithm was presented in [9] and applied in many areas: from marketing (predicting customer loyalty to a brand) and the criminal sphere (predicting homicide or relapse among parole), to credit scoring. Based on the financial reporting data, the authors [4] successfully use random forest models for forecasting defaults of companies from seven European countries (Finland, France, Germany, Italy, Portugal, Spain and the UK). In 1990, the authors [10] were among the first to use the neural network in predicting bankruptcies. A neural network was built with several hidden layers and using financial coefficients used in the Altman model as input. At the same time, the share of correctly classified companies was about 80%.

These algorithms often show higher efficiency, despite the fact that they are characterized by significant time and physical costs. Moreover, at present, there is a tendency in which algorithms based on one method are losing popularity, while ensemble or hybrid models are becoming more popular and demonstrate higher efficiency [11].

Since the 1970s, financial ratios derived from financial statements have been an important source for constructing default forecasting models. However, models based on accounting information are criticized because of the historical nature of the information used as input and not taking into account the volatility of the value of the company during the period analyzed. However, proponents of this approach argue that the inefficiency of capital markets can lead to more significant errors in predicting credit risks. In article [12], credit risk assessment models based on accounting and market information are compared. The authors conclude that the approaches considered do not have significant differences in the predictive ability, while these types of data are complementary, and the complex model shows the best result.

It can be concluded that market information can be a significant factor in predicting company insolvency. However, due to the fact that most of the companies examined do not have access to the stock market, financial statements become the only available source of information, and the use of market models becomes impossible.

## 2. Data description

The main data source in the work was the SPARK system (Interfax agency). Information about the default of companies was used in the “Unified Federal Register of Bankruptcy Information” database. In the study, the following companies were classified as construction companies (classification in SPARK):

✦ building;

✦ construction of engineering structures;

✦ specialized construction work (development and demolition of buildings, preparation of the construction site, finishing construction work).

An important issue is the definition of an insolvent company. In accordance with the Federal Law of October 26, 2002 No. 127-ФЗ (dated December 29, 2017) “On Insolvency (Bankruptcy)” [13], one sign of bankruptcy is considered to be a situation where the demands of creditors on monetary obligations are not fulfilled within three months from the date they were to be executed. The following definitions of bankruptcy are widespread in research: a company is not able to pay interest on a debt or part of its principal debt, the organization is monitored (a procedure that analyzes the financial situation and solvency of a debtor, as well as its ability to pay off debt), the company is not active for a long period of time, the company is in a state of liquidation. In our work, the category of bankrupts included companies in respect of which the legal bankruptcy procedure was launched, as well as companies that liquidated voluntarily. A similar classification is given in [4, 14]. It is noted that these companies are characterized by a critical financial situation and are often unable to fulfill their obligations.

Based on the financial statements, the values of fourteen coefficients reflecting the economic activity of the enterprise were calculated. At the same time, the following classification of financial indicators was proposed: profitability, liquidity, business activity, financial stability. A similar classification is given in [3]. Also included in the model are non-financial factors that reflect the size and age of the company. The variables are described in *Table 1.*

*Table 1.*

**Variable description**

| Group | Variables | Variable description |
|---|---|---|
| Dependent variable | Bankruptcy | 1 – if a default occurred in the next reporting period;<br>0 – otherwise |
| Profitability | Return on assets (ROA) | Net profit to assets ratio |
| | Return on equity (ROE) | Net profit to equity ratio |
| | Return on sales (ROS) | Net profit to revenue ratio |
| | Operating margin | Operating profit to revenue ratio |
| Liquidity | Current ratio | Current assets to current liabilities ratio |
| | Quick ratio | Receivables, financial investments and cash to current liabilities ratio |
| | Equity maneuverability ratio, net working capital (NWC) ratio | The difference between equity and non–current assets to equity ratio |
| Business activity | Accounts receivable (AR) turnover ratio | Revenue to receivables ratio |
| | Accounts payable (AP) turnover ratio | Cost of sales to accounts payable ratio |
| | Assets turnover ratio | Revenue to assets ratio |
| | Share of non–current assets | Non–current assets to total assets ratio |
| Financial stability | Autonomy ratio | Equity to assets ratio |
| | Share of retained earnings in revenue | Retained earnings to revenue ratio |
| | Interest coverage ratio (ICR) | Profit before tax and interest payable to interest payable ratio |
| Company size | Logarithm of company's assets | Assets logarithm |
| Age | Age | |

For the analysis of default events, the time range of 2011–2017 was chosen. The time horizon was divided into two blocks: a training sample (period 2011–2015) and a test sample (period 2016–2017). At the next stage, the following selection procedure was carried out:

1) removal of observations with missing data (for example, for which there is no information on the value of assets and revenue) or filling in gaps in the data, where possible;

2) removal of observations with obvious errors (for example, where the size of assets or the size of receivables is negative);

3) identification and removal of outlier observations, since their presence leads to biased results. The main algorithm used for this procedure is the three sigma rule.

As a result, 3981 organizations fell into the final sample, 390 of which defaulted. The training sample (period from 2011 to 2015) included 3300 construction companies, of which 325 defaulted. The test sample (the period from 2016 to 2017) included 681 companies, of which 65 defaulted. *Figure 1* shows the total number of companies and the number of companies that went bankrupt in the next reporting year, by year.

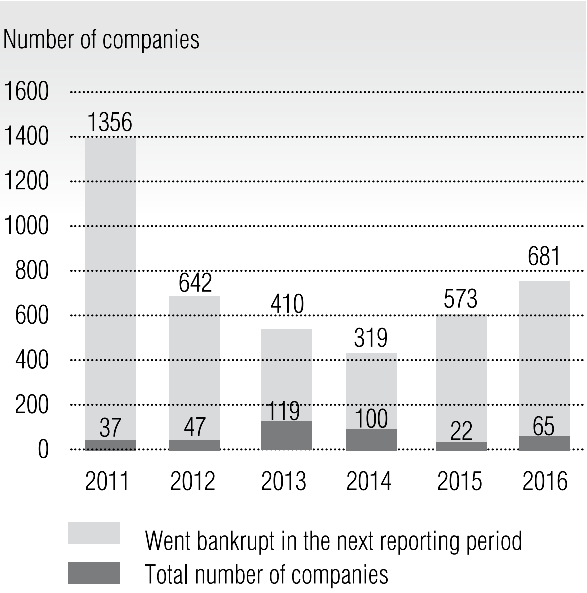


*Fig. 1.* The number of bankruptcies by year

The research data set is unbalanced (only 9.8% of companies defaulted). Therefore, when constructing the models, two techniques for working with unbalanced data were used: undersampling and oversampling.

Undersampling involves the use of input data containing all insolvent companies and a random selection of solvent companies. As a result, the proportion of insolvent to solvent companies increases. Also, when constructing such dependencies, it is recommended that this experiment be performed several times to obtain consistent results (in this study, the assumption is made that a consistent result is obtained after one experiment). Oversampling involves the use of input data containing all solvent companies, and "cloning" of insolvent companies until their number approaches the number of solvent companies. The search for the optimal share of the minority class in the training set is also the subject of research in this paper.

Cross-validations were used to find the optimal value of the share of insolvent companies in undersampling. When using oversampling, this approach is not recommended, since in this case, we see cloning of information which is used both in training and in testing the model (which leads to overfitting). To implement oversampling, the training set was divided into two subsets. The first (company default information for 2014) was used to test models, the second (remaining periods from 2011 to 2015) – to build models without cross-validation. The share of insolvent companies, in which the model is most important for the learning set, was used to compare models on the test set (2016–2017).

## 3. Description of models

Two parametric algorithms for constructing binary choice models were used in the work: logit and probit models with sampling corrections (using WOE) and without. These models are compared with algorithms based on machine learning methods (classification trees, random forests, artificial neural networks) which are described in the following sections.

Traditionally, the following metrics of model quality are distinguished: accuracy, sensitivity, specificity, area under the ROC curve, Gini coefficient, F-metric. The use of these metrics depends on the purpose of the analysis.

In this study, each of the models considered at the output has a range of values from 0 to 1; therefore, it is necessary to determine the cutoff threshold. The assignment of the cutoff threshold depends on the analyst's preferences regarding errors of the first and second kind, which leads to difficulties in comparing different models. Therefore, in this paper, the predictive power is estimated using ROC analysis, AUC (the area under the ROC curve), or Gini coefficient. The advantage of these metrics is that there is no need to determine the cutoff threshold and the ability to compare the quality of models regardless of the analyst's goals. The calculation of the Gini coefficient was carried out as follows:

$$Gini = (AUC - 0.5) \cdot 2 \cdot 100\%, \qquad (1)$$

where *AUC* is the area under the ROC curve.

Visual analysis of the effectiveness was carried out using the ROC curve. The greater the bend of the ROC curve, the higher the quality of the model, while the diagonal line corresponds to the complete indistinguishability of the two classes. Accordingly, the higher the value of the area under the ROC curve, the better the separation power of the model. Analysis of the ROC curve allows the user to select the ratio between sensitivity and specificity necessary for analysis. An example of constructing an ROC curve for one variable is presented in *Figure 2.*

In the work, as the calculation tool for econometric analysis and reflection of statistical conclusions, we used the programming language R, which is a free open source software environment.

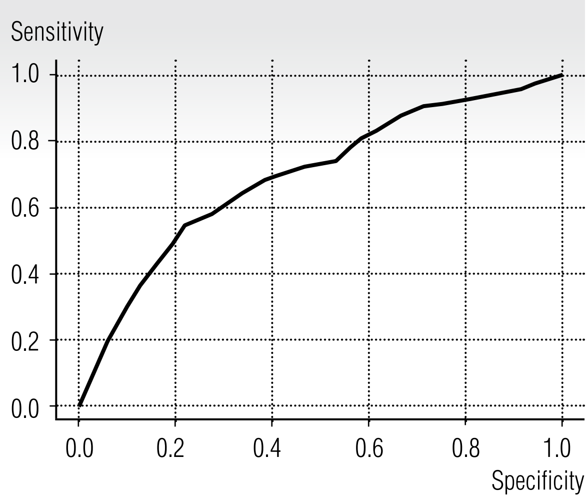


*Fig. 2.* ROC curve of the ROA factor

### 3.1. Binary selection models

Two algorithms for constructing binary selection models (logit and probit) were used in the work: without corrections for discretization and with corrections. Here is a description of the general algorithm (with discretization amendments) for constructing binary choice models.

**Step 1. Reduction factors to the discrete form.** In the process of solving a research question, most authors are faced with the problem of outliers. The given problem is no exception for this work. The traditional approach to solving it is the exclusion of such observations. However, the subjectivity of outlier determination and sample reduction are significant disadvantages of this approach. The paper uses the transition from discrete to continuous form, which leads to increased comparability of factors among themselves and the unity of approaches to assess the significance of factors. We carried out the quantile discretization procedure – replacing the initial values of factors with discrete values based on grouping by quantiles. The essence of this approach is as follows:

a) the values of the variables are ordered in ascending order;

b) the values of each indicator are divided into 10 groups (deciles are used);

c) the values in each group are replaced by points from 1 to 10 (the group with the lowest values gets 1 point, and the group with the highest values gets 10 points).

**Stage 2. Transformation of factors.** To convert factors, the WOE (weight of evidence) approach is used. The WOE indicator characterizes the degree of deviation of the level of defaults in this group from the average value in the sample. For each factor and for each group within the factors, it is necessary to calculate the number of companies in default and the number of companies not in default. The WOE for group i of a particular variable is calculated as follows:

$$WOE_i = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \tag{2}$$

where $d_i^{(1)}$ – the share of non-default companies belonging to group $i$ in the total number of non-default companies; $i = 1, 2, ..., k$; $k$ – number of variable categories;

$d_i^{(2)}$ – the share of companies in default owned by group $i$ in the total number of companies in default; $i = 1, 2, ..., k$; $k$ – number of variable categories.

In order to increase the linearity of variables and improve the accuracy of the model, all explanatory variables are replaced by WOE, which is a common technique in credit scoring [15].

**Stage 3. Assessment of the predictive power of factors.** After all values are converted to WOE, it is necessary to evaluate the importance of each factor. Two algorithms for assessing the significance of factors were used in the work: information value (information value, *IV*) and ROC analysis. The calculation of the value of information value (*IV*) is performed according to the following formula:

$$IV = \sum_{i=1}^{k}\left(d_i^{(1)} - d_i^{(2)}\right)\cdot WOE_i, \tag{3}$$

where $k$ – the number of categories of an independent variable (each factor has ten), the remaining notation – from formula (2).

Formula (3), which reflects calculation *IV*, is based on the summation of $WOE_j$, adjusted for the difference $\left(d_i^{(1)} - d_i^{(2)}\right)$. The main purpose of these calculations is to identify some indicator that reflects the ability of a variable to cluster some attribute. If this indicator is above 0.02, then the factor should be used in modeling [15].

The study applied the following criteria for selecting factors in the final model:

✦ acceptable quality of the model in accordance with the criterion of "information value" ($IV > 0.02$);

✦ the Gini coefficient in the one-factor model must be greater than 5%;

✦ economic assessment factor.

**Stage 4. The analysis of correlations.** When constructing a multi-factor model, factors with high correlation coefficients must be excluded. Correlation analysis avoids multicollinearity. Multicollinearity leads to model instability and increases standard deviations of factor estimates. The presence of multicollinearity is indicated by high values of pair correlation coefficients between the factors of variables. The criterion for determining high correlation may vary; for economic data, the threshold is usually set at 0.30–0.50. The criterion for high correlation in this model is a correlation coefficient greater than 0.5.

**Step 5: Multivariate analysis.** The modeling of the probability of the borrower's non-creditworthiness was carried out as follows:

$$\mathrm{P}\left(Y_i = 1|x_1, \ldots, x_n\right) = F\left(a_0 + a_1x_1 + \ldots + a_nx_n\right) = \\ = F\left(a_0 + \mathbf{x}'a\right). \tag{4}$$

In the case of the logit model, F (*) represents the logistic distribution function:

$$F\left(a_0 + \mathbf{x}'a\right) = \Lambda\left(a_0 + \mathbf{x}'a\right) = \frac{e^{a_0 + \mathbf{x}'a}}{1 + e^{a_0 + \mathbf{x}'a}}. \tag{5}$$

In the case of the probit model, F (*) is a normal distribution function:

$$F(a_0 + \mathbf{x}'a) = \Phi(a_0 + \mathbf{x}'a) = \int_{-\infty}^{a_0+\mathbf{x}'a} \varphi(\upsilon)d\upsilon, \qquad (6)$$

where $\varphi(\upsilon) = \frac{1}{\sqrt{2\pi}} e^{\left(-\upsilon^2/2\right)}$.

The calculation of the coefficients is carried out by the maximum likelihood method, which maximizes the probability of the joint implementation of events (solvency and insolvency). The standard error of the coefficients was estimated with a Newey–West correction for heteroskedasticity and first-order autocorrelation.

**Step 6. Model specification.** To select the optimal combination of factors, the Backward Selection method was used – sequential exclusion of factors (i.e. insignificant variables are sequentially excluded from the model, which includes all factors selected in the one-factor analysis). At the same time, the level of statistical significance is tested using p-value calculated according to the results of logistic regression. As a result, factors with p-value less than 10% were selected.

**Stage 7. Validation of the model.** Choosing the best model. The problem of overfitting requires a model validation procedure. This problem is manifested in the fact that the "trained" model has good results on the training sample, but does not give accurate forecasts for the test sample. To solve this problem, two approaches were used.

The first approach is the "mixing algorithm", the idea of which is as follows:

1. 80% of the companies from the training set are randomly selected;

2. The coefficients of the model are estimated;

3. It is evaluated whether signs are preserved at coefficients, and whether the factors considered are significant;

4. Steps 1–3 are repeated 1000 times; the stability of the signs is checked.

Based on the results obtained, it can be concluded whether the signs of the coefficients for all variables are stable, and how the sign of the coefficient depends on the initial sample.

The second approach is ROC analysis. Analysis of the values of AUC and Gini on the test set helps to make a conclusion about the quality of the models obtained.

### 3.2. Machine learning models

Logistic analysis (as well as probit analysis) are traditional popular tools for predicting bankruptcies, but they have a number of disadvantages associated with low predictive power, the presence of restrictions on use. Therefore, at the moment, machine learning algorithms have become widespread.

**Classification trees.** Today classification trees are the foundation for building more complex machine learning algorithms, such as random forests and boosting algorithms (GBM, XGBoost). In this paper, the CART algorithm (classification and regression trees) is considered. A distinctive feature of this algorithm is that it provides only two possible options for the development of the event, which is suitable for realizing the purpose of this study. The main idea of CART is to split the primary set into two subsets so that the bankrupt companies are in one set, and solvent organizations are in the other. The difficulty in using this method is to determine the moment of stopping the "splitting of sets," since the problem of overfitting arises. The following stopping rules are distinguished:

✦ the measure of "purity" is less than a certain value;

✦ restriction on the number of nodes or layers of a tree;

✦ size of the parent node;

✦ size of the descendant node.

The rules themselves are set using cross validation. Despite the fact that there are a number of examples of the successful use of this method in forecasting defaults [5], this method has several disadvantages: high sensitivity to input data, susceptibility to overfitting and the difficulty of determining the optimal tree architects.

**Random forests.** Random forests appeared as a modification of decision trees and, accordingly, often provide more accurate predictive results. Random forests consist of a user-defined number of classification trees that are generated using a modified CART algorithm. The scheme of this algorithm is presented in *Figure 3.* Two approaches were used in the algorithm: each tree is trained on its own subsample of initial data (bootstrapped data); different subsets of factors are used in constructing classification trees. These actions lead to the construction, and then to the "voting of trees" regarding the belonging of an object to a certain class.

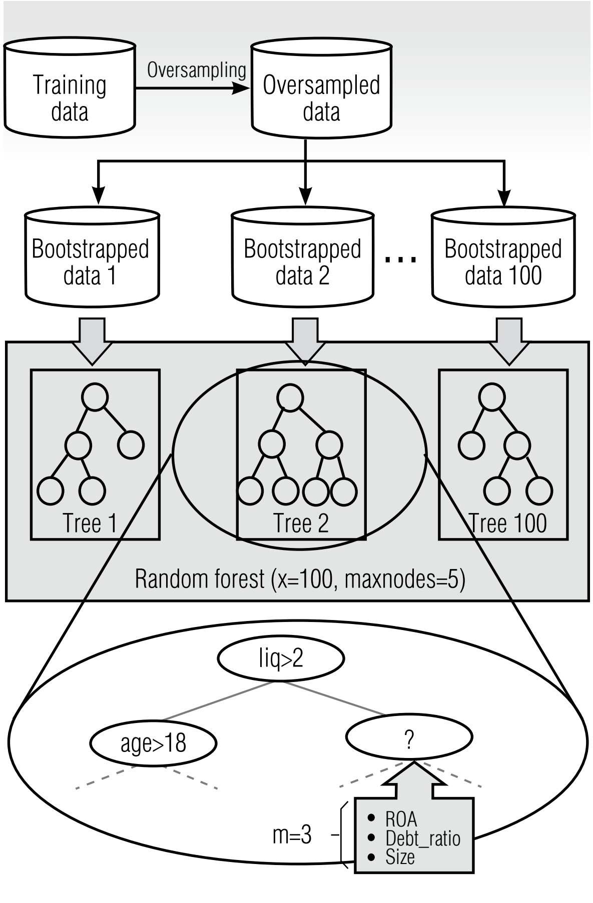


*Fig. 3.* Random forest algorithm [4]

Unlike regression models, which are quite sensitive to outliers, random forest (RF) is more robust to this problem. The advantage of random forest is higher efficiency in case of imbalance of data (which is relevant for our task), as well as less exposure to overfitting. The disadvantage of the algorithm is less transparency (in contrast to classification trees) and, accordingly, lower interpretation. There is relative difficulty in the process of determining the parameters of a random forest. The determination of the parameters (number of trees, number of factors used in building one tree of factors, maximum number of nodes in one tree) was carried out using cross-validation.

Random forests are often used to determine the significance of a variable. The idea of assessing the importance of a factor is based on the fact that a permutation of the values of an important variable should lead to a significant increase in the error rate on the test set.

Artificial neural networks. Currently, neural network modeling is gaining popularity, especially when predicting phenomena with uniform attributes. Using a set of input parameters, the network architecture is selected. When in the simplest version it is represented by three layers. The first layer contains nodes (neurons) for input variables (each neuron has only one input from the external environment). The second layer contains an arbitrary number of "hidden" neurons and is therefore called a hidden layer. The third layer contains neurons that are responsible for the result. Moreover, in the tasks of forecasting bankruptcies, the last layer contains only one neuron. Between the input and hidden neurons, a connection with certain weights is set. For example, for the *j*-th neuron in the intermediate layer and the input data, the following linear dependence will be determined:

$$a_j = \sum_{i=1}^{N} \omega_{ij} \cdot x_i \, , \qquad (7)$$

where $a_j$ – the value of the $j$-th neuron;

$\omega_{ij}$ – weight of $j$-th neuron with variable $x_i$.

Each value $a_j$ is converted using some activation function to obtain the actual resulting value $z_j$ of neuron $j$. Since our study predicts two classes, it is convenient to use the logistic function as an activation function:

$$z_j = f(a_j) = \frac{1}{1+e^{-a_j}}, \qquad (8)$$

where $z_j$ – normalized value of the $j$-th neuron;

$a_j$ – the value of the $j$-th neuron.

A similar procedure is carried out for subsequent layers. The $z_j$ values are again weighted, and then converted using the activation function to obtain the result in the final layer. Many minimization methods are distinguished. Their idea is that, starting from the initial value of weight $\omega^0$, a sequence of vectors of weight coefficients $\omega^1$, $\omega^2$, ..., $\omega^k$ is generated, such that with each iteration of the algorithm the value of the function of the quality criterion decreases:

$$E\left(\omega^{k+1}\right) < E\left(\omega^{k}\right), \qquad (9)$$

where $\omega^k$ – weight value after the $k$-th step of training;

$\omega^{k+1}$ – weight value after the $k$+1-th step of training.

*Figure 4* shows an example of one of the resulting neural network models.

One of the most common methods used to train neural network models is the steepest descent method. In this algorithm, the adjustment of the weights is performed in the direction of the maximum reduction of the quality criterion, i.e. in the opposite direction to the gradient vector. Despite the fact that the steepest descent method converges to the optimum value of $\omega^*$ rather slowly, it is a common method of finding the minimum in many statistical libraries.

One of the difficulties of training a neural network is that the quality criterion function can have many local minima. As a result, after the initialization of the model, one can come to a local minimum, which will negatively affect the results obtained on the test set. To overcome this problem, weights are randomly sorted, and the learning algorithm itself is repeated several times. The optimal parameters (the number of neurons in the inner layer, the number of inner layers), as well as for classification trees and random forests, were determined using cross-validation.

To increase the efficiency of the neural network, as well as speed up the learning process, preliminary data processing is necessary. A simple and efficient preprocessing step involves scaling and centering data.

## 4. Comparative analysis of the models

In accordance with the classification used, the models considered showed good quality. *Table 2* shows the sorting of the best models in the group in descending order of model quality. The best quality was shown by an artificial neural network with one hidden layer and four neurons using the oversampling algorithm. It is reflected that the use of a logistic model with sampling and transition to WOE leads to a significant increase in the accuracy of models (the Gini coefficient increases on average by 15%). It is noteworthy that the quality of the models corresponds to the AUC level in such works [4, 5, 11].

The results of one-way analysis using regression models indicate that all the factors considered can be used to build binary choice models, since for each of them the AUC value for univariate analysis is higher than 0.5. The final multivariate model included eight factors out of sixteen factors reflecting different aspects of the risks of construction companies: liquidity (current ratio, equity ratio), profitability (return on equity), solvency (interest coverage ratio), turnover (asset turnover), business activity (share of non-current assets), non-financial predic-

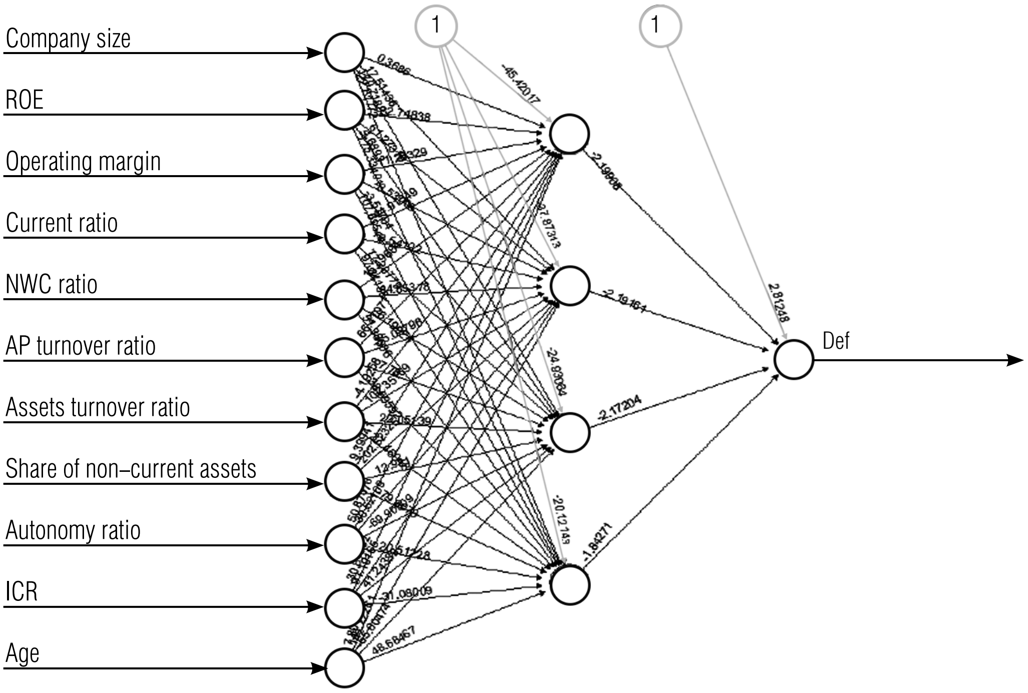


*Fig. 4.* Example of neural network architecture

*Table 2.*

**Quality assessment of models on a test set**

| No | Model | Gini coefficient, % | The share of insolvent companies on the test set, % |
|---|---|---|---|
| 1 | Artificial neural networks (oversample) | 59.6 | 50 |
| 2 | Artificial neural networks | 58.9 | 9.8 |
| 4 | Logit model (oversample) | 57.9 | 25 |
| 3 | Probit model (oversample) | 57.6 | 20 |
| 5 | Logit model | 57.6 | 20 |
| 6 | Logit model (undersample) | 57.3 | 20 |
| 7 | Artificial neural networks (undersample) | 56.0 | 50 |
| 8 | Random forests (undersample) | 52.4 | 15 |
| 9 | Random forests | 50.6 | 9.8 |
| 10 | Random forests (oversample) | 48.7 | 10 |
| 11 | Classification trees (oversample) | 45.0 | 15 |
| 12 | Log model without discretization | 42.2 | 9.8 |
| 13 | Classification trees (with penalties for incorrect classification of a minority class) | 40.0 | 9.8 |
| 14 | Classification trees | 38.0 | 9.8 |
| 15 | Classification trees (undersample) | 38.0 | 50 |

tors (age and size of the company). The inclusion of these factors leads to an increase in the efficiency of traditional models (the Gini coefficient increases from 0.38 to 0.58). Moreover, these models showed resistance to overfitting. In a multivariate analysis of the hypothesis regarding the sign of the dependence of the probability of insolvency on the alleged regressors were confirmed. Significant differences in accuracy indicators between the logit and probit models were not found.

This conclusion is consistent with many works, since the logistic distribution function and the distribution function of the standard normal random variable behave approximately the same, and the differences are associated with more “heavy tails” of the logistic distribution function.

Due to the stability of nonparametric algorithms to multicollinearity, all the factors considered earlier were used to build models based on machine learning methods. Analysis of classification trees and random forests showed that among the most influential factors were the coefficient of maneuverability of equity and the coefficient of autonomy (the largest drop in the Gini index in the random forest algorithm, the first partition in classification trees). This means that if a company has a significant amount of debt burden and it shows a negative financial result (in the balance sheet its equity is negative), this is an important indicator of the company’s insolvency in the next reporting period. At the same time, non-financial factors (age, company size) turned out to be practically insignificant, which is reflected in *Figure 5*. Thus, the large size and long life of the company in the market cannot guarantee stability in the Russian market.

The dynamics of the average value of the Gini coefficient depending on the share of insolvent companies with optimal parameters on the training set using undersampling and oversampling (*Figure 6*) shows that in the forecasting problem with this data structure, the influence of the share of insolvent companies on the training set does not significantly affect the forecast potential one or another method. This conclusion is consistent with the work of Demeshev and Tikhonova [14]. The negative dynamics of the quality metric with an increase in the share of insolvent companies indicates a bias in the construction of the algorithm (for example, in the “random forest using oversampling” algorithm).

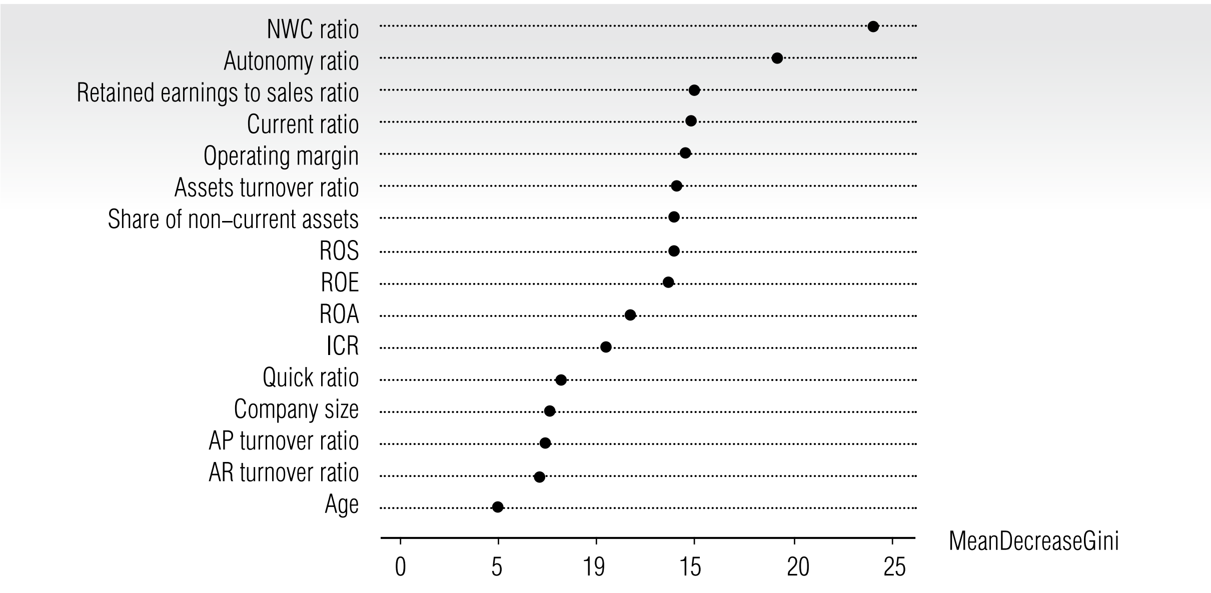


*Fig. 5*. Determination of the most significant parameters. Random forest algorithm

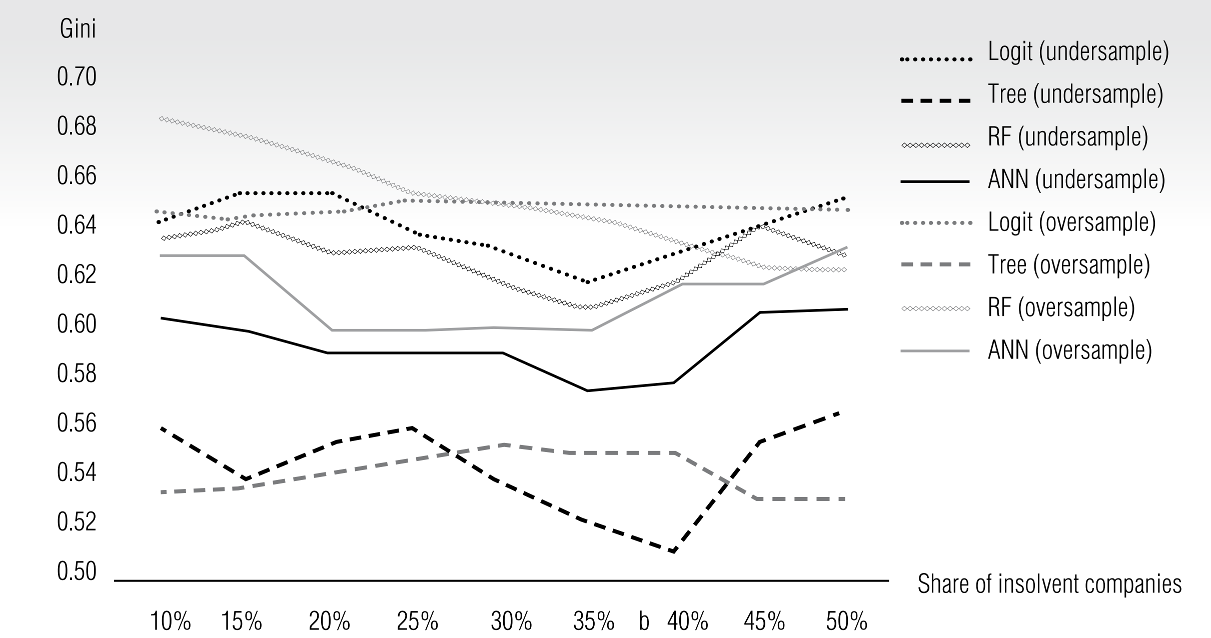


*Fig. 6.* The average value of the Gini coefficient on the training set

The use of the method of combating imbalance depends on the type of model used. For logistic regression, artificial neural networks and classification trees, oversampling has shown higher quality. However, the use of oversampling in the random forest method leads to overfitting. Therefore, for random forests, undersampling is more effective.

## Conclusion

The use of a particular model depends on the goal of the analyst. In forecasting problems, nonlinear algorithms, as a rule, show a higher result. Therefore, the use of neural networks and random forests is more acceptable for this type of task. However, these models lose to the binary choice models in costs (time, computational) for calculations, as well as in interpretation.

The algorithms we examined showed acceptable quality for use in the tasks of forecasting bankruptcies of construction companies. As expected, the best model was an artificial neural network. Traditional sampling models have shown good results, while their results can be easily interpreted, and the calculation time is minimal. Despite the advantages of classification trees (ease of interpretation, the absence of restrictions on the type of variables, the absence of the need to specify the relationship form in an explicit form), this algorithm showed instability and low accuracy of predictions.

In the future, it seems promising to include other nonlinear algorithms in comparison, for example, models based on boosting (GBM, XGBoost), support vector models, etc. Moreover, in this work, the category of bankrupts includes companies in respect of which the legal bankruptcy procedure has begun, as well as companies that have liquidated voluntarily. In the future, it seems possible to distinguish between these categories using a single federal register of bankruptcy information and identify companies in respect of which the legal bankruptcy procedure has begun. It also seems possible to conduct an intersectoral comparison of the methods considered, determine the maximum forecasting horizon at which signs of bankruptcy appear, diversify within individual industries and use macroeconomic variables in modeling. ■

## References

1. Karminsky A.M., Kostrov A.V., Murzenkov T.N. (2012) *Approaches to evaluating the default probabilities of Russian banks with econometric methods.* Working paper WP7/2012/04 (Series "Mathematical methods of decision analysis in economics, business and policies"). Moscow: HSE (in Russian).
2. Kostrov A.V. (2016) Comparison of statistical classification methods to predict Russian banks failures. *Management of Financial Risks*, vol. 47, no 3, pp. 162–180 (in Russian).
3. Tserng P., Chen P.-C., Huang W.-H., Lei M.C., Tran Q.H. (2014) Prediction of default probability for construction firms using the logit model. *Journal of Civil Engineering and Management*, vol. 20, no 2, pp. 247–255. DOI: 10.3846/13923730.2013.801886.
4. Behr A., Weinblat J. (2017) Default patterns in seven EU countries: A random forest approach. *International Journal of the Economics of Business*, vol. 24. No 2. pp. 181–222. DOI: 10.1080/13571516.2016.1252532.
5. Gepp A., Kumar K. (2015) Predicting financial distress: A comparison of survival analysis and decision tree techniques. *Procedia Computer Science*, vol. 54, pp. 396–404.
6. Tam K.Y., Kiang M.Y. (1992) Managerial applications of neural networks: the case of bank failure predictions. *Management Science*, vol. 38, no 7, pp. 926–947. DOI: 10.1287/mnsc.38.7.926.
7. Bogdanova T.K., Shevgunov T.Ya., Uvarova O.M. (2013) Using neural networks for predicting solvency of Russian companies on manufacturing industries. *Business Informatics,* no 2, pp. 40–48 (in Russian).
8. Breiman L., Friedman J., Olshen R., Stone C. (1984) *Classification and regression trees.* Wadsworth, New York: Chapman and Hall.
9. Breiman L. (2001) Random forests. *Machine Learning,* vol. 45, no 1, pp. 5–32.
10. Odom M.D., Sharda R. (1990) A neural network model for bankruptcy prediction. Proceedings of *International Joint Conference on Neural Networks. San Diego, USA, 17–21 June 1990*, vol. 2, pp. 163–168. DOI: 10.1109/IJCNN.1990.137710.
11. Kumar P.R., Ravi V. (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques: A review. *European Journal of Operational Research*, vol. 180, no 1, pp. 1–28. DOI: 10.1016/j.ejor.2006.08.043.
12. Trujillo-Ponce A., Samaniego-Medina R., Cardone-Riportella C. (2014) Examining what best explains corporate credit risk: accounting-based versus market-based models. *Journal of Business Economics and Management*, vol. 15, no 2, pp. 253–276. DOI: 10.3846/16111699.2012.720598.
13. The Federal Law of 26 October 2002, No 127-FZ (as amended on 29 December 2017) *"On insolvency (bankruptcy)"* (in Russian).
14. Demeshev B.B., Tikhonova A.S. (2014) *Default prediction for Russian companies: intersectoral comparison.* Working paper WP2/2013/05 (Series WP2 "Quantitative Analysis of Russian Economy"). Moscow: HSE (in Russian).
15. Siddiqi N. (2012) *Credit risk scorecards: developing and implementing intelligent credit scoring.* John Wiley & Sons.

## About the authors

**Alexander M. Karminsky**

Dr. Sci. (Econ.), Dr. Sci. (Tech.);

Professor, Department of Finance, National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: karminsky@mail.ru

ORCID: 0000-0001-8943-4611

**Roman N. Burekhin**

Doctoral Student, Doctoral School on Economics, National Research University Higher School of Economics,
20, Myasnitskaya Street, Moscow 101000, Russia;

E-mail: romanvia93@yandex.ru

ORCID: 0000-0003-1130-0175