National Research University Higher School of Economics

As a manuscript

Afanasev Vladislav Viktorovich

USING NON-FINANCIAL INFORMATION FOR DEFAULT PREDICTION IN SERVICES SECTOR

PhD Dissertation Summary
for the purpose of obtaining an academic degree
Doctor of Philosophy in Economics

Academic supervisor:
Candidate of Economic Sciences
Ichkitidze Yuri Rolandovich

JEL: G21, G32, G33, C10, C50

Moscow – 2024

Problem statement and motivation

The services sector is the one of the major sectors in the economy of Russian Federation. According to World Bank data[1], services formed 54% of Russian GDP in 2022. Today, there is no single definition of the services sector. In this study, firms in the following industries are classified as service sector:

- Tourism, accommodation and passenger transportation services;
- Food services;
- Education;
- Medical and social services;
- Cultural, sports and entertainment services;
- Housing and utilities services;
- Financial services;
- Other services (personal services, veterinary services, repair services, etc.).

The need for accurate default prediction is especially relevant for the services sector. Firstly, the services sector firms form a big share of bankruptcy cases (around 10% as stated in Fedresurs statistics for 2022), and, having on average less assets to be disposed in case of bankruptcy, are riskier for the creditors in terms of potential debt repayment. The share of debts, which is paid to the creditors during insolvency procedures in the services sector is very low, for example, it was only 3,4%, in 2019. Thus, default prediction for services sector firms may be of interest for the credit organizations and the players on the services market.

One of the ways to handle uncertainties regarding defaults of companies and reduce credit risks for lenders is the use of probabilistic default prediction models. Traditionally, such models are based on financial data calculated for some time before the default. The researchers have proposed different

[1] World Bank data. Services, value added (% of GDP). URL: https://data.worldbank.org/indicator/NV.SRV.TOTL.ZS Accessed 10.08.2023

probabilistic models, which are based on the analysis of financial ratios of the firms. Most of them used statistical techniques, such as MDA (Altman, 1968), Logistic Regression (Gruszczyński, 2004; Hunter and Isachenkova, 2001; Ohlson, 1980; Sirirattanaphonkun and Pattarathammas, 2012), or Machine Learning algorithms (Altman *et al.*, 1994; Cao *et al.*, 2020; Coats and Fant, 1993; Odom and Sharda, 1990; Ravi Kumar and Ravi, 2007; Zhang *et al.*, 1999). Such models are still commonly used for credit risk estimation.

The majority of the prior studies report the predictive accuracy of 80% and more (Altman, 1968; Sirirattanaphonkun and Pattarathammas, 2012). However, there are some preconditions, which allow to claim that the existing instruments of credit risk estimation may be inappropriate for service firms in Russia.

According to the estimate, provided by World Economics[2], about 38% of Russian GDP is formed by shadow economy. It means that the financial ratios for some of the firms may be biased if the firms combine formal and informal activity. Also, a lot of businesses in services sector can be presented as groups of entities, including sole proprietorships (with no reporting), limited liability companies (on different tax regimes) etc. Also, fraudulent actions by the management or shareholders may occur. Finally, a very small number of firms are audited, which can potentially lead to the mistakes in the reporting. Thus, the financial data obtained for legal entities may not reflect the real condition of the business, so the traditional approach to default prediction may give poor results.

One way to improve the quality of default prediction is to level out the effect of industry peculiar properties and develop industry specific models to develop industry specific models like those proposed by several studies (Costa *et al.*, 2022; Kanapickienė *et al.*, 2023; Psillaki *et al.*, 2010; Situm, 2023; Давыдова and Беликов, 1999). This study covers only services sector.

[2] World Economics data. URL: https://www.worldeconomics.com/Informal-Economy/

But, more important, there is a way to capture for bad predictive ability of the financial data - include non-financial variables.

Some researchers attempted to explore this domain, (Altman *et al.*, 2010, 2016; Blanco-Oliver *et al.*, 2016) and found out that the usage of non-financial variables, such as, for example, value of legal claims or negative audit reports can lead to more accurate default prediction. Thus, using non-financial information can be beneficial in terms of forecast quality.

To sum up, it seems that the traditional approach to default prediction (using financial data) may result in low accuracy for Russian firms, thus it may be beneficial to use non-financial information for making high-quality default prediction in the services sector. Thus, the research questions raised in the study, are:

RQ1: "How significant can be the improvement in default prediction quality in the services sector if one uses non-financial data?"

RQ2: "What are the non-financial factors that significantly improve the quality of default prediction in the services sector?"

Object and subject of the study

The *object* of this study is the models for predicting default among private firms operating in the service sectors in Russia.

As the *subject*, the study examines the potential for increasing the accuracy of default prediction by incorporating non-financial data alongside financial ratios as explanatory factors.

Purpose and objectives of the study

The purpose of this study is to increase the accuracy of insolvency prediction models for Russian private firms in the service sector by using non-financial data in alongside financial ratios, which serves for facilitating investment decisions by the counterparties of such firms and credit organizations.

The objectives of the study include:

- Research the existing approaches to default prediction in Russian and foreign studies, including the conventional approach (using financial ratios for default prediction) and the new approach - using non-financial data for this purpose;
- Confirm or refute the low ability of financial ratios to predict the future insolvency of Russian private firms in the service sector compared to firms from developed western markets;
- Develop default prediction models for selected industries from the services sector and find the evidence for the improvement in the quality of prediction if non-financial data are used along with financial ratios, or refute it;
- Reveal the non-financial variables with the highest impact on the default prediction accuracy;
- Formalize the approach for data collection, applicable for default prediction studies, from the perspective of the theoretical forecast date estimation (the date, at which the values of the independent variables should be calculated);
- Develop an approach to assess the quality of the models, based on the analysis of predicted probabilities of default for every observation.

Brief literature review

The origins of default prediction modeling trace back to seminal studies by (Beaver, 1966) and (Altman, 1968). Beaver pioneered credit risk assessment by employing specific financial ratios to estimate bankruptcy likelihood, while Altman developed the Z-Score model using discriminant analysis to predict defaults in US-listed firms. Despite Beaver's innovative use of financial ratios, his method lacked a single indicator for default probability, leading to interpretive challenges. Altman's Z-Score model categorized firms into three default risk groups based on five financial determinants, demonstrating a predictive accuracy

of 95% over 50 years. However, criticism of the Z-Score model includes potential bias arising from correlated independent variables (Евстропов, 2008) and its inapplicability to small and medium-sized firms due to the use of market capitalization .

James Ohlson's 1980 work (Ohlson, 1980) introduced logit regression analysis for default prediction, identifying six determinants of default for American SMEs. These included various financial ratios complemented by dummy variables indicating recent losses and unfavorable liabilities-to-assets ratios. Other pioneer studies include (Springate, 1978), (Zmijewski, 1984), (Blum, 1974), and (Deakin, 1972), which addressed sampling biases and early warning signs of default.

The conventional approach to default prediction primarily utilizes financial ratios as explanatory variables. From early models to recent works (D'Amato and Mastrolia, 2022; Jandaghi et al., 2021; Zhao and Lin, 2023), financial data covering profitability, liquidity, capital structure, and turnover are employed to predict default (Jaki and Ćwięk, 2021). A study by (du Jardin, 2008) found that 93% of default prediction studies utilized financial ratios, indicating a wealth of knowledge in this domain. And it seems that the quality metrics reported in various default prediction studies are high (up to 98%).

However, the conventional approach may not be suitable for Russian service firms due to issues such as business disaggregation, shadow operations, fraud and mistakes in reporting. One proposed solution involves incorporating non-financial data. In recent academic research, integrating non-financial data into default prediction models has emerged as a promising trend aimed at improving accuracy. For example, (Altman *et al.*, 2010) note this shift, highlighting a significant departure from traditional approaches.

A wide array of non-financial variables can be utilized in default prediction models, the choice of variables is only limited by practical reasoning. For example, (Blanco-Oliver et al., 2016; Karminsky and Burekhin, 2019;

Lugovskaya, 2010) use age and size of the observed firms as the default predictor. The logic underlying the use of these data is that smaller firms are hypothesized to be more risky because of less assets to dispose, and younger firms are hypothesized to be more risky due to the same reason and less experience on the market. These are probably the most used non-financial factors in such kind of studies. (Bhimani et al., 2013) report a significant increase in area under ROC curve when using macroeconomic, management, ownership and financial support related variables along with financial ratios, hypothesizing that financial support, good quality of management and favourable macroeconomic conditions should decrease the credit risk. Ownership variables such as ownership concentration or management ownership are also used by other researchers (e.g. Ragab and Saleh, 2021). (Altman et al., 2010; Kanapickienė et al., 2023), use the information about the late submission of financial reporting among other variables, hypothesizing that a firm which overdue the financial reporting should be more risky. (Blanco-Oliver et al., 2016) use legal claims data as the predictor, hypothesizing that a firm, which faces more legal claims, should have more debts to pay. Some researchers (e.g. Makeeva and Sinilshchikova, 2020; Stevenson *et al.*, 2021) use textual data analysis to predict insolvency including, for example, the tone of the loan application text (Filomeni et al., 2024). Another example is the use of financial reporting quality measures as predictors, e. g. as shown in (Costa *et al.*, 2022).

The utilization of non-financial data in default prediction is not yet widespread, but promising findings suggest significant improvement in prediction accuracy. For instance, (Altman et al., 2010) reported an 8% increase in area under the ROC curve when incorporating non-financial variables, (Lugovskaya, 2010) reported 16% increase in the accuracy metric, (Bhimani et al., 2013) reported a 21% increase in the area under the ROC curve etc. This highlights the potential of non-financial variables to complement financial indicators and enhance the effectiveness of risk assessment practices.

Methodological basis of the study

The empirical study consists of the following steps, necessary to answer the research questions:

1. Testing the hypothesis of poor applicability of the conventional approach to default prediction (using financial ratios only as predictors) to Russian services sector by comparing the accuracy of the models built for Russian firms and their peers from developed European markets.
2. Tests for the use of non-financial data for the purpose of default prediction for selected service industries: auto repair industry, healthcare industry, housing and utilities management industry, microfinance industry to assess the effect of non-financial data inclusion and reveal the most valuable variables.

Econometric tools utilized in the study

The study utilizes three econometric tools - logistic regression, K-Nearest Neighbors (KNN), and random forest to perform default prediction. Logistic regression, a commonly used linear classification method in default prediction research, offers interpretability by assessing the contribution of each independent variable. It constrains predicted outcomes between 0 and 1, estimating "probabilities" of default. Unlike ordinary linear regression, logistic regression's key assumption is a linear relationship between independent variables and the log odds of the dependent variable.

The logistic function transforms the output of a linear regression model, estimating the probability of default based on the linear coefficients assigned to independent variables. The model categorizes observations into default or non-default groups using a threshold probability typically set at 50%. Maximum likelihood estimation optimizes coefficients to maximize the likelihood function, often incorporating L1 regularization to control the number of variables.

KNN, a straightforward classification algorithm, classifies observations

based on similarity to neighboring data points. The appropriate number of neighbors (k) is crucial for accuracy and is set to the square root of total observations in this study. Euclidean distance measures similarity, so the data should be normalized before modeling. The classification process involves identifying the k most similar firms from the training dataset and assigning the observation to one of the classes based on a majority vote among these "neighbors".

Random forest, chosen for its high predictive accuracy and variable importance assessment capabilities, is an ensemble of classification trees. Each tree predicts default, and the most frequent class among trees becomes the final prediction of the model. The trees are trained on randomly picked subsets (with replacement) from the training data, ensuring randomness. The core aim of a classification tree is to split the training data to minimize Gini impurity, ensuring each leaf contains only one class of observations or a dominant class share.

In summary, logistic regression offers interpretability, KNN provides simplicity, and random forest delivers high predictive accuracy and variable importance assessment. These econometric tools offer distinct approaches to default prediction, making them valuable assets for financial analysis and risk management.

The evaluation of binary classification model quality relies on standard metrics derived from the confusion matrix. The matrix facilitates the assessment of model performance through three key indicators: accuracy, sensitivity (recall), and specificity. Accuracy reflects the proportion of correct predictions relative to the total number of observations. Sensitivity measures the proportion of true positive predictions in the total number of actual positives (default cases), while specificity measures the proportion of true negative predictions in the total number of actual negatives (non-default cases).

In the study, algorithms estimate default probabilities, which are then converted into binary outcomes using a chosen cutoff point. The selection of this

cutoff point significantly impacts model performance metrics such as sensitivity, accuracy, and specificity. To visualize this dependency, a receiver operator curve (ROC) is employed, illustrating sensitivity against 1-specificity across various cutoff values.

The ideal classifier exhibits sensitivity and specificity values of 1 for respective cutoff values below and above 50%. A classifier with zero predictive ability yields a diagonal line on the ROC curve. The area under the ROC curve (AUC) serves as a quality metric, with a value of 1 indicating an ideal classifier and 0.5 signifying a random chance classifier.

These methodologies ensure robust model development and facilitate accurate default prediction, crucial for effective risk management in financial analysis.

Data used in the study

The study implies creating classification models for firms, thus, the unit of observation is a firm, the dependent variable is a binary variable (1 if default, 0 otherwise), the independent variables are financial and non-financial attributes of the firms, available on the theoretical forecast date.

The financial attributes differ across analyzed industries, but always cover 4 types of financial ratios: liquidity, profitability, turnover and solvency ratios.

The non-financial factors also differ across analyzed industries and include information related to legal claims against the firm, inspections, tenders, key changes in the firm (location, CEO, shareholders), age of the firm etc.

The data used in this study was collected by the author from three major sources of information: Amadeus database for private firms in EU, SPARK-Interfax database for Russian private firms (financial data, data related to the age of firms and legal claims against the firms), “Всероссийский Бизнес Центр” website (non-financial information).

The main findings of the study and the provisions defended

1. *It has been found that the accuracy of default prediction for Russian service sector firms using only financial data is lower compared to service sector firms from developed European markets, and this difference is statistically significant.*

Firstly, the prediction accuracy of Logit Regression, K-Nearest Neighbors, and Random Forest classification algorithms when trained on a dataset of Russian service firms was compared with those trained on a dataset of service firms from developed European markets. The hypothesis driving this comparison was that the accuracy of prediction would likely be lower for Russian service firms due to potential biases in financial reporting, such as shadow operations and business disaggregation, compared to their peers in developed European markets.

During the Soviet period, in the absence of a legislative framework supporting private business, the provision of services to the population was monopolized by state structures. However, there was a shortage of everyday services, which led to the emergence of the informal sector. Traces of this historical influence remain in the contemporary landscape, as also shown by surveys of entrepreneurs (Williams *et al.*, 2013). According to a study by the Forum for the Study of Eastern Europe and Emerging Economies (Putniņš and Sauka, 2020), the size of the shadow economy in Russia is estimated at almost 45% of GDP. If firms combine “white” and shadow activities, the reporting of such firms may not reflect the real condition of these firms. In a situation of business disaggregation, which is most often present due to tax optimization (Качалин, 2011; Трошкова and Ильясов, 2023), there is a need to consolidate financial statements in order to accurately assess the overall state of the firms. However, this task is often impossible due to the lack of group reporting. Moreover, some parts of the corporate structure may be represented by sole proprietorships or companies using simplified tax regimes, which exempts them from providing detailed financial statements. Thus, analysts have to rely on data

for a single legal entity to evaluate the financial performance, which entails potential bias. Fraud and reporting errors (according to the Ministry of Finance, only about 2% of legal entities in Russia are audited) can also distort financial performance.

The dataset contains information on firms from various industries within the services sector, including tourism, accommodation, passenger transportation, dining & catering, education, medical & social services, culture, sports & entertainment services, and other services (personal services, veterinary services, repair services). These firms are classified using the OKVED-2 classification for Russian firms and the NACE Classification for European firms.

Two datasets were prepared: one for Russian service firms that faced default between 2017 and 2020 and one for service firms from developed European Union economies. Each defaulting firm was paired with a healthy peer firm, matched based on total assets. The dependent variable is a dummy variable indicating default (1) or non-default (0), while independent variables are financial ratios calculated based on the reporting available on the default date.

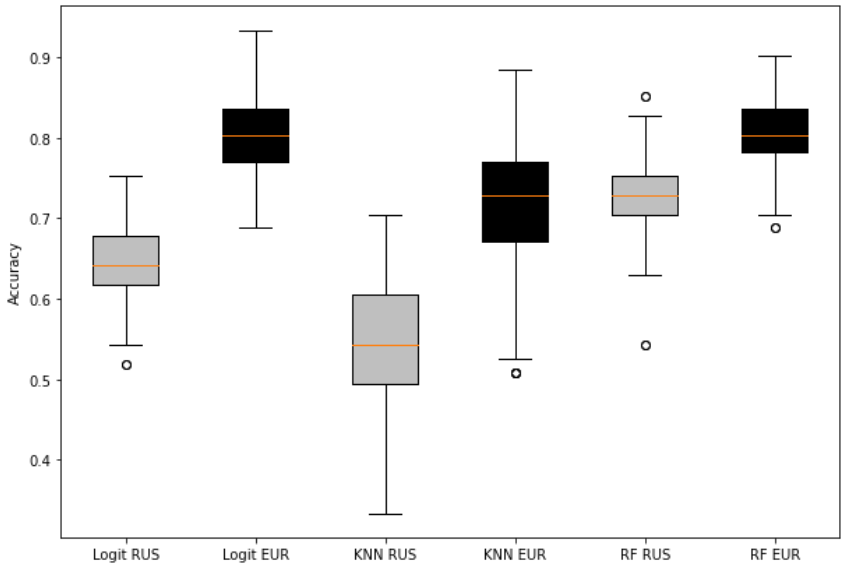
The control dataset contains information on defaults and healthy firms from developed European Union countries. The choice of this control group is based on the data, showing that issues of shadow operations and business disaggregation are less prevalent in these countries compared to Russia.

The datasets were then divided into training and test sets 100 times, with classification algorithms trained on 100 training sets and applied to 100 test sets. Then the mean prediction accuracies were compared. The main hypothesis is tested using the Mann-Whitney test to compare the mean accuracy across test sets for Russian and European service firms. Three machine learning algorithms - Logistic Regression, K-Nearest-Neighbors, and Random Forest were employed to ensure robustness in the analysis.

The results of the classification analysis using Logit, KNN, and Random Forest algorithms demonstrated notable differences in prediction accuracy

between Russian and European service firms. The results suggest that the prediction accuracy of all three classification algorithms is considerably lower for Russian service firms compared to their European counterparts. Figure 1 serves to contribute to this statement.

Figure 1. Classification results



Source: Prepared by the author

2. *Insolvency prediction models were constructed for four service sectors, for which similar studies had not previously been conducted using Russian data. It was found that the resulting models show high accuracy.*

Secondly, using several industries from services sector default prediction models were built. Accuracy tests conducted on these models indicate their potential utility in evaluating the credit risk of firms operating within their industries. Decent accuracy was received for every of the industries used in the study. The major classification metrics for testing data are presented in Table 1.

According to Kazakov and Kolyshkin (Казаков and Колышкин, 2018), the majority of bankruptcy prediction models in Russia do not demonstrate accuracy higher than 70%, thus, the accuracy of the models developed in this study can be assessed as high.

Table 1. Classification results (best performing models[3] are shown)

	Auto repair services	**Healthcare services**	**Housing and utilities management services**	**Microfinance services**
# of observations	2240	138	1134	135
% of defaulted	9% (1:10)	33% (1:2)	33% (1:2)	20% (1:4)
Accuracy	70,7%	95,5%	84,8%	78,0%
Sensitivity	68,0%	77,8%	83,3%	62,5%
Specificity	73,6%	100%	85,6%	81,8%
AUC	0,82	0,98	0,91	0,80

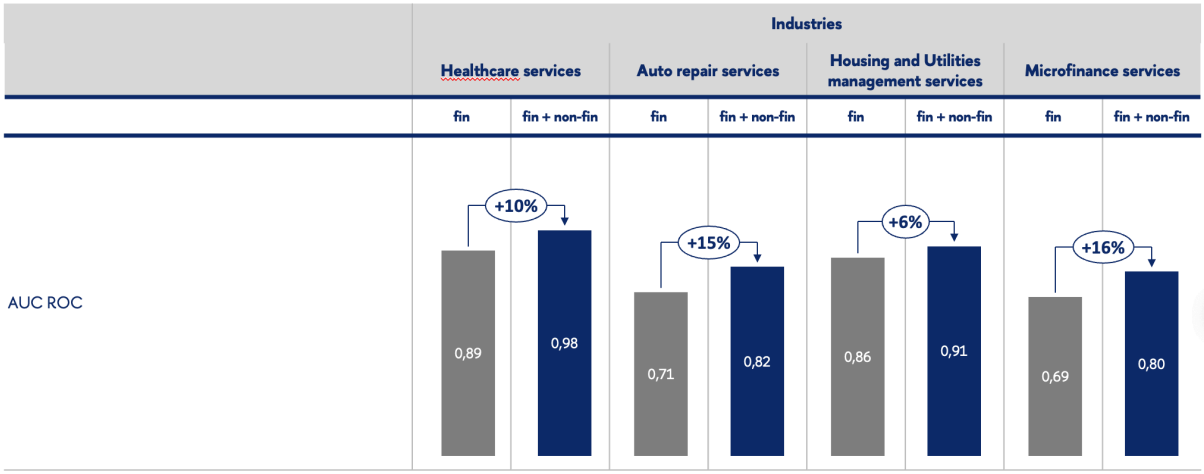
Source: Prepared by the author

[3] The table shows the results of the models with the highest AUC among the models evaluated and tested on a sample of companies from relevant industries.

3. *A positive impact of non-financial data on the accuracy of the models was found. Inclusion of these variables increases the ROC AUC indicator by an average of 9 percentage points.*

In all the industries considered the inclusion of non-financial data leads to higher predictive ability of the models. The figure below shows the area under ROC curve for every of the considered industries firstly for the model based on financial data only, then for the model, based on financial and non-financial data together. Best performing specifications were chosen.

Figure 2. The effect of non-financial data inclusion



Source: Prepared by the author

There were several non-financial variables utilized in the study. Some of them were used by several researchers before. One of the example is the data related to legal claims against the firms, which was previously used by (Blanco-Oliver *et al.*, 2016). This data, expressed in the number of legal claims and the sum of these claims (divided by the assets of the firm), was found to be valuable in terms of the accuracy of prediction.

4. *New non-financial variables, which **had not** been previously utilized by researchers in this field (such as data on business inspections, participation in tenders, and key changes in the company), were found. These variables has a positive effect on the accuracy of the models.*

The same conclusion can be made regarding the additional non-financial data, which was used in this study for the first time ever. For example, the information related to inspections faced by the firms turned to be a good default predictor for housing and utilities management firms, the number of tenders won turned to be a good default predictor for auto repair firms, number of changes in CEO turned to be a good default predictor for microfinance firms. This underscores the importance of considering a broader range of data sources for more accurate estimation of credit risk.

The list of all non-financial variables used and information on their significance in the models for different industries is given in Table 2. The table shows only those variables that were not filtered out at the stage of the test for significance of the difference in the mean values for defaulted and non-defaulted companies and were included in the model for at least one industry. A dash (-) in the table means that this variable was not included in the model due to insignificant difference in means for defaulted and non-defaulted firms, or was not used for this industry. The blocks, starting with "LR" in the table indicate the significance of the variables in logistic regression models in the format "significant/insignificant (*significance level*)". If a variable was included in several models within the same industry, the best indicator among these models is placed in the table. Results for those variables that were significant at the 10% level and below are shown in bold. The blocks, starting with "RF" show the importances of the variables in the random forest models in the format "*rank of the variable in the ranking by importance* out of *the number of variables in the model*". If a variable was included in several models within the same industry, the best indicator among these models is placed in the table. The results for those

variables that were included in the top 50 percent in terms of importance are shown in bold.

Table 2. Non-financial variables used and their significance

Variables	Auto repair services	Healthcare services	Housing and utilities management services	Microfinance services
Number of inspections against the firm	LR: insignificant RF: 13 out of 13 by importance	-	-	-
Number of violations identified during the inspections	-	LR: insignificant RF: 9 out of 9 by importance	LR: insignificant	-
Share of inspections with violations identified	-	-	**LR: significant (0,1%)**	-
Number of tenders won	**LR: significant (5%)** RF: 12 out of 13 by importance	-	-	-
Number of changes in shareholders or CEO	LR: insignificant RF: 11 out of 13 by importance	-	-	**LR: significant (10%)** RF: 7 out of 8 by importance
Number of location changes	-	-	-	**LR: significant (10%)** RF: 7 out of 8 by importance
Number of notifications about the address incorrectness	-	-	-	LR: insignificant **RF: 1 out of 8 by importance**
Number of legal claims filed against the firm in the arbitrage court	**LR: significant (0,1%)** **RF: 3 out of 13 by importance**	**LR: significant (0,1%)** **RF: 3 out of 9 by importance**	**LR: significant (0,1%)**	**LR: significant (10%)** **RF: 4 out of 8 by importance**
Sum of legal claims filed against the firm (in rubles) per 1 ruble of assets	LR: insignificant **RF: 1 out of 13 by importance**	LR: insignificant **RF: 1 out of 9 by importance**	**LR: significant (0,1%)**	LR: insignificant RF: 5 out of 8 by importance

Source: prepared by the author

5. *It has been found that models containing both financial and non-financial data as predictors estimated using a non-linear machine learning algorithm (random forest) allow for more accurate insolvency prediction. However, at the same time, the linear algorithm (logistic regression) provides sufficient accuracy in some cases and can be used at least to assess the contribution and the direction of influence of each explanatory factor.*

In each case when both logistic regression models and random forest models were built on the same data, the models estimated using random forest algorithm demonstrated relatively higher accuracy on the test sample. The increase in accuracy compared to the models based on logistic regression is, on

average, 16 percentage points in terms of ROC AUC (see Table 3).

Table 3. The comparison of accuracy metrics for models estimated using logistic regression and random forest algorithm

Accuracy metric	Auto repair services		Healthcare services		Microfinance services	
	Logit	Random forest	Logit	Random forest	Logit	Random forest
Accuracy	73,8%	70,7%	75,0%	95,5%	61,0%	78,0%
Sensitivity	62,7%	68,0%	78,9%	77,8%	75,0%	62,5%
Specificity	75,0%	73,6%	66,7%	100%	57,6%	81,8%
AUC	0,69	0,82	0,81	0,98	0,62	0,80

Source: prepared by the author

It can be seen that at least in some cases, models estimated using logistic regression algorithm demonstrate relatively high accuracy (for example, in the auto repair services industry or in private clinics industry). Therefore, given that logistic regression allows for assessing the contribution of each variable, and more important, the direction of effect (positive or negative), it is recommended to use this algorithm when constructing classification models, including when nonlinear machine learning algorithms are used.

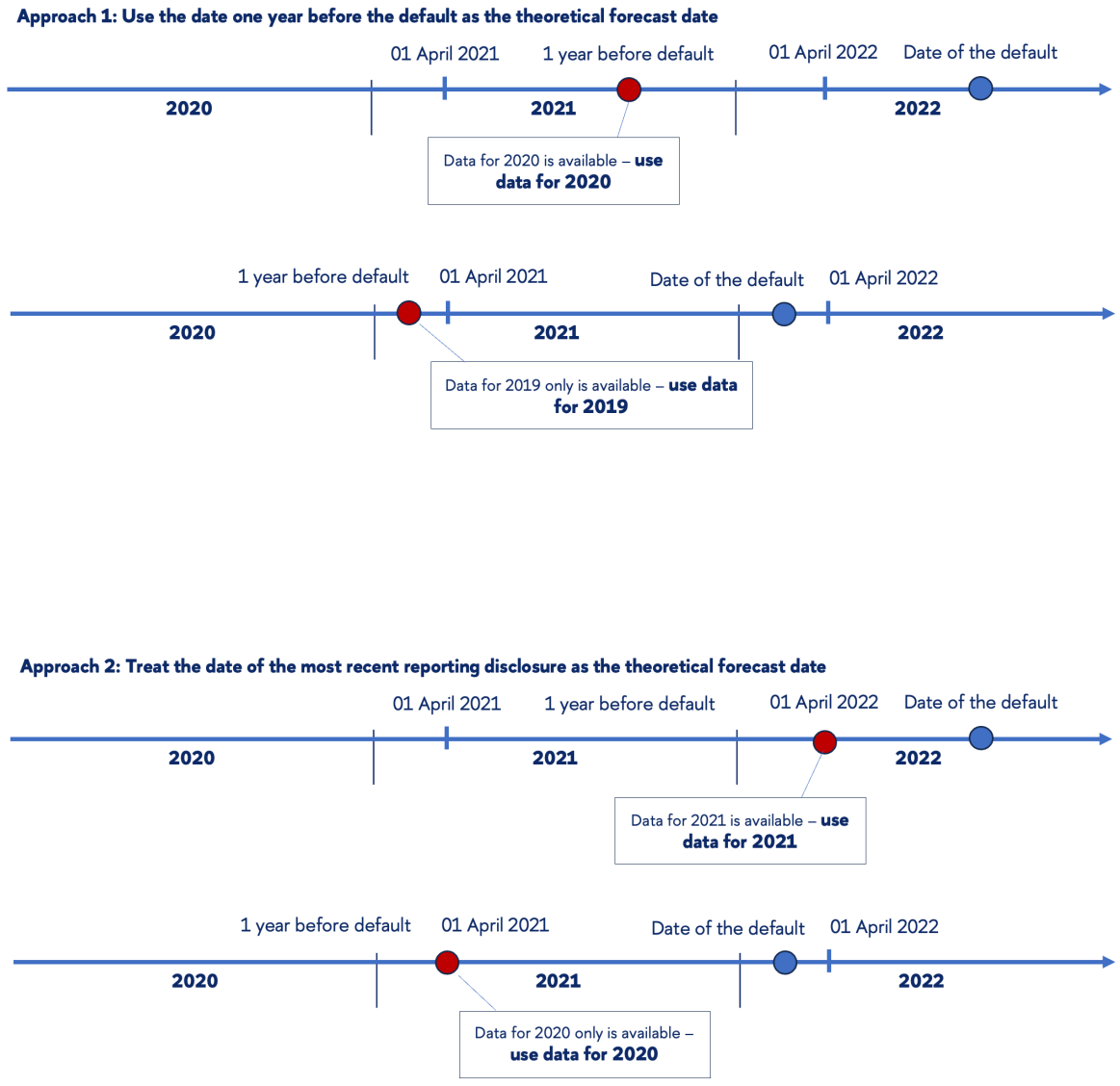
6. *An approach for assessing the theoretical forecast date (the date on which values of independent variables should be calculated), which allows to control for the data availability at the moment of implied forecasting, is developed and utilized in the study.*

The study proposes two approaches for determining the time period used to calculate independent variables in default prediction models. The first approach, termed the "Year-before" approach, involves modeling default prediction one year prior to the actual default date. This allows for predictions to be made a year in advance, with the theoretical forecast date set exactly one year before the default occurs. In contrast, the second approach, referred to as the

"Available reporting" approach, utilizes the most recent financial reporting available on the default date to calculate financial variables. Many of the studies, related to default prediction do not state the time period, for which the values of the predictors are calculated, do not pay attention to checking if the data used in the model was actually accessible at the theoretical forecast date. This study provides a ready-to-use framework to collect the data properly.

The approach is shown in Figure 3.

Figure 3. Two approaches used for estimating the theoretical forecast date.

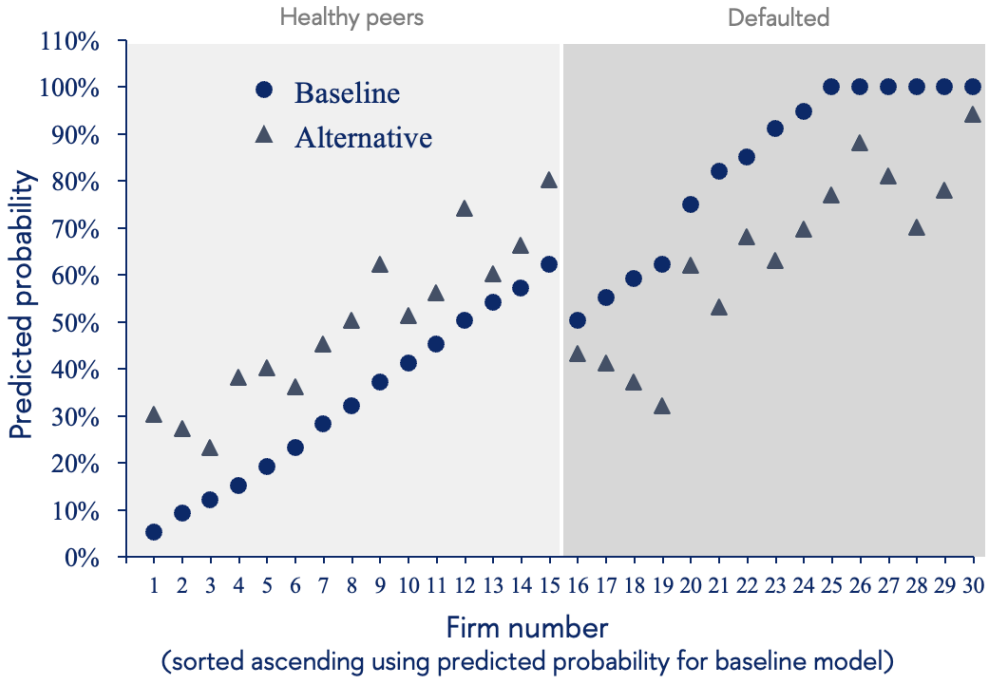


Source: prepared by the author

7. *The study introduces an approach to assessing model quality based on the analysis of predicted default probabilities for each observation offering the advantage of simplicity in application, suitable for practitioners without special knowledge.*

The proposed method for evaluating the models' quality involves examining the predicted probabilities of default. Figure 4 illustrates an example of plotting the predicted probabilities for the testing dataset, comparing a baseline model to some alternative model. The x-axis represents the firm IDs, while the y-axis displays the predicted probabilities. The graph delineates two distinct areas: the left segment corresponds to data points for healthy peers (no default), whereas the right segment corresponds to defaulted firms. Firms are sorted based on the probabilities predicted by the baseline model.

Figure 4. An example of predicted probabilities plot



Source: prepared by the author

Then, four hypotheses can be tested using a test for equality of means to judge about the quality of the models and to compare them:

H1. The mean probabilities predicted by the base model for defaulting

firms do not statistically differ from the mean probabilities predicted by the base model for non-defaulting firms (the base model does not perform adequately).

H2. The mean probabilities predicted by the base model for defaulting firms do not statistically differ from the mean probabilities predicted by the alternative model for defaulting firms (both models provide similar results for defaulting firms).

H3. The mean probabilities predicted by the base model for non-defaulting firms do not statistically differ from the mean probabilities predicted by the alternative model for non-defaulting firms (both models provide similar results for non-defaulting firms).

H4. The mean probabilities predicted by the alternative model for defaulting firms do not statistically differ from the mean probabilities predicted by the alternative model for non-defaulting firms (the alternative model does not perform adequately).

Using, for example, Mann-Whitney test one can test these hypotheses and judge about the quality of the classifiers.

Theoretical contribution

Default prediction models were developed and tested using financial and non-financial data as explanatory variables. The influence of including non-financial variables was analyzed, and a positive impact on the prediction accuracy has been concluded. Non-financial variables that are valuable in terms of the accuracy of insolvency prediction models have been identified.

Managerial implications

The study also has the potential managerial implication. The credit organizations may be interested in including non-financial data in their original credit scoring models. The service firms themselves, their counterparties and

employees can benefit from this study by discovering, which non-financial variables can be the indicators of the future insolvency of the service firms.

Contribution to Methodology

A formalized approach to assessing the theoretical forecast date was formulated and an algorithm for collecting data was presented. These may be useful in conducting future research on this topic. The study also proposed a method for evaluating model quality that allows visual evaluation of model quality and comparison between models. The advantage of this approach lies in its ease of use in practice and accessibility for any user without specialized training or advanced knowledge in the field of machine learning and classification models. Also, the approaches to modelling were compared and it was found that models based on non-linear machine learning algorithm (random forest) better fit the data, than those based on linear algorithm (logit).

The scientific novelty of this study

1. It is the first study covering the comparison of the default prediction accuracy between Russian services firms and services firms from developed EU markets. Previous studies, firstly, are focused mainly on the foreign countries – there are few studies related to Russia. Secondly, the existing research does not cover specifically service sector in Russia. Finally, the previous studies focus on default prediction for one country, not comparing the accuracy across economies.
2. It is the first study to cover the topic of the default prediction of housing & utilities management firms and auto repair firms. These industries were not examined in the existing literature in terms of default prediction, however, being industries with high market volume, big number of small players, thus, high need for accurate credit risk estimation.

3. The impact of non-financial data was assessed in Russian context. New variables not utilized in the existing literature but having impact on the accuracy of default prediction are suggested in the study (e.g. inspections data, government procurement data).
4. This study formalizes the approach to estimating the theoretical forecast date (the date, at which the values of the independent variables should be calculated). Most of the previous studies do not pay much attention to it or the approaches are not clear.
5. This study suggests an approach to assess the quality of the models, based on the analysis of predicted probabilities of default for every observation.

Structure of the work

The thesis consists of an introduction, 4 chapters, a conclusion, a list of references and appendices. The total volume of work is 137 pages of the main text (not including references and appendix), the bibliography includes 140 titles.

Approval of scientific results

The results of this study were presented in the form of presentation on the following conferences / research seminars:

1. Modern Econometrics Tools and Applications. Nizhniy Novgorod, 23-25 September 2021. Presentation: “Can the conventional approach to bankruptcy prediction be applicable to Russian service firms?”.
2. Analytics for Management and Economics Conference (AMEC Junior). Saint-Petersburg, 28 May 2022. Presentation: “Default Prediction for Housing and Utilities Services Management Firms Using Non-financial Data”.
3. Russian Economic Congress. Ekaterinburg, 11-15 September 2023. Presentation: “Default prediction for auto repair firms using non-financial data”
4. Research Seminar on Empirical Studies of Banking. Moscow, 20 March 2024. Presentation: “Using non-financial information for default prediction in services sector”

List of author's original articles on the topic of the dissertation research

1. Afanasev V. Default Prediction Model for Emerging Capital Market Service Companies // Journal of Corporate Finance Research. 2023. Vol. 17. No. 1. P. 64-77. (Afanasev, 2023)
2. Afanasev V., Tarasova J. Default Prediction for Housing and Utilities Management Firms Using Non-Financial Data // Научно-исследовательский финансовый институт. Финансовый журнал. 2022. Vol. 14. No. 6. P. 91-110. (Afanasev and Tarasova, 2022)
3. Egor O. Bukharin, Mangileva S. I., Afanasev V. Default Prediction for Russian Food Service Firms: Contribution of Non-Financial Factors and Machine Learning // Journal of Applied Economic Research. 2024. Vol. 23. No. 1. P. 206-226. (Bukharin *et al.*, 2024)

References

1. Afanasev, V. (2023), "Default Prediction Model for Emerging Capital Market Service Companies", Journal of Corporate Finance Research, Vol. 17 No. 1, pp. 64–77, doi: 10.17323/j.jcfr.2073-0438.17.1.2023.64-77.
2. Afanasev, V. and Tarasova, Y. (2022), "Default Prediction for Housing and Utilities Management Firms Using Non-Financial Data", Financial Journal, Vol. 14 No. 6, pp. 91–110, doi: 10.31107/2075-1990-2022-6-91-110.
3. Altman, E. (1968), "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", The Journal of Finance, Vol. 23 No. 4, pp. 589–609.
4. Altman, E., Iwanicz-Drozdowska, M., Laitinen, E. and Suvas, A. (2016), "Financial and nonfinancial variables as long-horizon predictors of bankruptcy", The Journal of Credit Risk, Vol. 12, pp. 49–78, doi: 10.21314/JCR.2016.216.
5. Altman, E., Sabato, G. and Wilson, N. (2010), "The value of non-financial information in SME risk management", The Journal of Credit Risk, Vol. 6, pp. 95–127, doi: 10.21314/JCR.2010.110.
6. Altman, E.I., Marco, G. and Varetto, F. (1994), "Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience)", Journal of Banking & Finance, Vol. 18 No. 3, pp. 505–529, doi: 10.1016/0378-4266(94)90007-8.
7. Beaver, W.H. (1966), "Financial Ratios As Predictors of Failure", Journal of Accounting Research, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 4, pp. 71–111, doi: 10.2307/2490171.
8. Bhimani, A., Gulamhussen, M.A. and Lopes, S. da R. (2013), "The Role of Financial, Macroeconomic, and Non-financial Information in Bank Loan Default Timing Prediction", European Accounting Review, Vol. 22 No. 4, pp. 739–763, doi: 10.1080/09638180.2013.770967.
9. Blanco-Oliver, A., Diéguez, A., Alfonso, M.D. and Vazquez Cueto, M. (2016), "Hybrid model using logit and nonparametric methods for predicting micro-entity failure", Investment Management and Financial Innovations, Vol. 13, pp. 35–46, doi: 10.21511/imfi.13(3).2016.03.
10. Blum, M. (1974), "Failing Company Discriminant Analysis", Journal of Accounting Research, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 12 No. 1, pp. 1–25, doi: 10.2307/2490525.
11. Bukharin, E.O., Mangileva, S.I. and Afanasev, V.V. (2024), "Default Prediction for Russian Food Service Firms: Contribution of Non-Financial Factors and Machine Learning", Journal of Applied Economic Research, Vol. 23 No. 1, pp. 206–226, doi: 10.15826/vestnik.2024.23.1.009.
12. Cao, Y., Liu, X., Zhai, J. and Hua, S. (2020), "A two-stage Bayesian network model for corporate bankruptcy prediction", International Journal of Finance & Economics, Vol. 27 No. 1, pp. 455-472, doi: https://doi.org/10.1002/ijfe.2162.

13. Coats, P.K. and Fant, L.F. (1993), “Recognizing Financial Distress Patterns Using a Neural Network Tool”, Financial Management, Financial Management Association International, Vol. 22 No. 3, pp. 142-155.
14. Costa, M., Lisboa, I. and Gameiro, A. (2022), “Is the Financial Report Quality Important in the Default Prediction? SME Portuguese Construction Sector Evidence”, Risks, Multidisciplinary Digital Publishing Institute, Vol. 10 No. 5, pp. 1-24, doi: 10.3390/risks10050098.
15. D’Amato, A. and Mastrolia, E. (2022), “Linear discriminant analysis and logistic regression for default probability prediction: the case of an Italian local bank”, International Journal of Managerial and Financial Accounting, Inderscience Publishers Ltd., Vol. 14 No. 4, pp. 323-343, doi: 10.1504/IJMFA.2022.126552.
16. Deakin, E.B. (1972), “A Discriminant Analysis of Predictors of Business Failure”, Journal of Accounting Research, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 10 No. 1, pp. 167–179, doi: 10.2307/2490225.
17. Filomeni, S., Bose, U., Megaritis, A. and Triantafyllou, A. (2024), “Can market information outperform hard and soft information in predicting corporate defaults?”, International Journal of Finance & Economics, Vol. 29 No. 3, pp. 3567–3592, doi: 10.1002/ijfe.2840.
18. Gruszczyński, M. (2004), “Financial Distress of Companies in Poland”, Department of Applied Econometrics, Warsaw School of Economics, Working Papers, No. 1-04, doi: 10.2139/ssrn.902256.
19. Hunter, J. and Isachenkova, N. (2001), “Failure risk: A comparative study of UK and Russian firms”, Journal of Policy Modeling, Vol. 23 No. 5, pp. 511–521, doi: 10.1016/S0161-8938(01)00064-3.
20. Jaki, A. and Ćwięk, W. (2021), “Bankruptcy Prediction Models Based on Value Measures”, Journal of Risk and Financial Management, Multidisciplinary Digital Publishing Institute, Vol. 14 No. 1, pp. 1-14, doi: 10.3390/jrfm14010006.
21. Jandaghi, G., Saranj, A., Rajaei, R., Ghasemi, A. and Tehrani, R. (2021), “Identification of the Most Critical Factors in Bankruptcy Prediction and Credit Classification of Companies”, Iranian Journal of Management Studies, University of Tehran, Vol. 14 No. 4, pp. 817–834, doi: 10.22059/ijms.2021.285398.673712.
22. du Jardin, P. (2009), “Bankruptcy prediction models: How to choose the most relevant variables?”, Bankers, Markets & Investors, pp. 39–46.
23. Kanapickienė, R., Kanapickas, T. and Neciunas, A. (2023), “Bankruptcy Prediction for Micro and Small Enterprises Using Financial, Non-Financial, Business Sector and Macroeconomic Variables: The Case of the Lithuanian Construction Sector”, Risks, Multidisciplinary Digital Publishing Institute, Vol. 11 No. 5, pp. 1-33, doi: 10.3390/risks11050097.
24. Karminsky, A. and Burekhin, R. (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.
25. Lugovskaya, L. (2010), “Predicting default of Russian SMEs on the basis of financial and non-financial variables”, Journal of Financial Services Marketing, Vol. 14 No. 4, pp. 301–313, doi: 10.1057/fsm.2009.28.
26. Makeeva, E. and Sinilshchikova, M. (2020), “News Sentiment in Bankruptcy Prediction Models: Evidence from Russian Retail Companies”, Корпоративные Финансы | ISSN: 2073-0438, Vol. 14 No. 4, pp. 7–18, doi: 10.17323/j.jcfr.2073-0438.14.4.2020.7-18.
27. Odom, M. and Sharda, R. (1990), “A Neural Network Model for Bankruptcy Prediction”, Vol. 2, pp. 163–168 vol.2, doi: 10.1109/IJCNN.1990.137710.
28. Ohlson, J.A. (1980), “Financial Ratios and the Probabilistic Prediction of Bankruptcy”, Journal of Accounting Research, [Accounting Research Center, Booth School of Business, University of Chicago, Wiley], Vol. 18 No. 1, pp. 109–131, doi: 10.2307/2490395.
29. Psillaki, M., Tsolas, I.E. and Margaritis, D. (2010), “Evaluation of credit risk based on firm performance”, European Journal of Operational Research, Elsevier, Vol. 201 No. 3, pp. 873–881.
30. Putniņš, T. and Sauka, A. (2020), The Shadow Economy in Russia: New Estimates and Comparisons with Nearby Countries.
31. Ragab, Y.M. and Saleh, M.A. (2021), “Non-financial variables related to governance and financial distress prediction in SMEs–evidence from Egypt”, Journal of Applied Accounting Research, Emerald Publishing Limited, Vol. 23 No. 3, pp. 604–627, doi: 10.1108/JAAR-02-2021-0025.
32. Ravi Kumar, P. and 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.
33. Sirirattanaphonkun, W. and Pattarathammas, S. (2012), “Default Prediction for Small-Medium Enterprises in Emerging Market: Evidence from Thailand”, Seoul Journal of Business, College of Business Administration, Vol. 18 No. 2, pp. 25-54

34. Situm, M. (2023), “Financial distress in the Austrian tourism industry: hotels and restaurants analysis”, European Journal of Tourism Research, Vol. 34, doi: 10.54055/ejtr.v34i.2992.
35. Springate, G.L.V. (1978), Predicting the Possibility of Failure in a Canadian Firm: A Discriminant Analysis, Simon Fraser University.
36. Stevenson, M., Mues, C. and Bravo, C. (2021), “The value of text for small business default prediction: A Deep Learning approach”, European Journal of Operational Research, Vol. 295 No. 2, pp. 758–771, doi: 10.1016/j.ejor.2021.03.008.
37. Williams, C.C., Nadin, S., Newton, S., Rodgers, P. and Windebank, J. (2013), “Explaining off-the-books entrepreneurship: a critical evaluation of competing perspectives”, International Entrepreneurship and Management Journal, Vol. 9 No. 3, pp. 447–463, doi: 10.1007/s11365-011-0185-0.
38. Zhang, G., Hu, M.Y., Patuwo, B.E. and Indro, D.C. (1999), “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis”, European Journal of Operational Research, Vol. 116 No. 1, pp. 16-32.
39. Zhao, Y. and Lin, D. (2023), “Prediction of Micro- and Small-Sized Enterprise Default Risk Based on a Logistic Model: Evidence from a Bank of China”, Sustainability, Multidisciplinary Digital Publishing Institute, Vol. 15 No. 5, pp. 1-13, doi: 10.3390/su15054097.
40. Zmijewski, M.E. (1984), “Methodological Issues Related to the Estimation of Financial Distress Prediction Models”, Journal of Accounting Research, Vol. 22, pp. 59-82, doi: 10.2307/2490859.
41. Давыдова, Г.В. and Беликов, А. (1999), “Методика количественной оценки риска банкротства предприятий”, Управление Риском, Vol. 23, No. 3, pp. 13-20.
42. Евстропов, М. (2008), “Оценка возможностей прогнозирования банкротства предприятий в России”, Вестник Оренбургского Государственного Университета, Федеральное государственное бюджетное образовательное учреждение высшего образования «Оренбургский государственный университет», Россия, Оренбург, No. 4, pp. 25–32.
43. Казаков, А. and Колышкин, А. (2018), “Разработка моделей прогнозирования банкротства в современных российских условиях”, Вестник Санкт-Петербургского Университета. Экономика, Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский государственный университет», Россия, Санкт-Петербург, No. 2, pp. 241–266.
44. Качалин, Д.С. (2011), “Анализ российских моделей дробления (реорганизации) бизнеса, обеспечивающих соответствие его масштабов специальным режимам налогообложения”, Финансовая Аналитика: Проблемы и Решения, Vol. 47, No. 5.
45. Трошкова, С. and Ильясов, Д. (2023), “Дробление бизнеса: понятие и налоговые последствия”, Инновационная экономика и общество, Vol. 40 No. 2, pp. 53–60.