DOI: https://doi.org/ 10.17323/j.jcfr.2073-0438.17.1.2023.64-77
JEL classification: G33, G21, C51

Default Prediction Model for Emerging Capital Market Service Companies

Vladislav Afanasev
Postgraduate Student, Lecturer,
School of Economics and Management, Department of Finance, National Research University Higher School of Economics, St. Petersburg, Russia,
vvafanasev@hse.ru, ORCID

Abstract

The author tested the hypothesis that default prediction based on financial data may be inapplicable to Russian service sector organizations by analyzing the differences in the accuracy of models based solely on financial data for service providers from Russia and developed European countries.

Logistic regression, Random Forest and K-nearest neighbors machine learning methods were used as modeling tools on a sample of 404 Russian firms and 304 firms from developed European countries.

The results suggest that the prediction error is significantly higher in the case of Russian firms than in the case of firms from the control group (European service firms). Thus, the use of financial ratios for default prediction for service firms in Russia seems insufficient.

These findings can be used by organizations that provide credit scoring, and by any other market participants interested in the financial stability assessment of their counterparties.

Keywords: service sector, default prediction, credit risk, machine learning algorithms

For citation: Afanasev, V. Default Prediction Model for Emerging Capital Market Service Companies. *Journal of Corporate Finance Research.* 2023;17(1): 64-77. https://doi.org/ 10.17323/j.jcfr.2073-0438.17.1.2023.64-77

Introduction

The conventional approach to default prediction implies using financial ratios as determinants of defaults. Since the late 1960s numerous researchers have demonstrated that financial ratios are good default predictors, starting with the famous paper of Edward Altman [1] and ending with some recent papers of both foreign [2] and Russian researchers [3; 4].

Over the course of these 60 years, default prediction using financial ratios has developed along with the advancement of statistical techniques, underlying it. Simple linear classification algorithms, like Multiple Discriminant Analysis [1], Logistic Regression [5–8] or Probit Regression [9; 10] are now partly substituted with more precise non-linear Machine Learning algorithms [11–17].

The set of financial ratios used as default predictors has also expanded. The researchers have added non-trivial predictors, such as the growth rate of income [18] or the standard deviation of stock returns [19]. Some researchers also prove that non-financial predictors can improve prediction accuracy [20–26]. However, there are still very few papers that deal with non-financial predictors of default in general and related to Russian firms in particular. One possible explanation of this fact could be the high predictive power of conventional default prediction models (based on financial ratios).

At the same time, using only financial ratios for default prediction seems to be inefficient in case of developing economies, and namely in case of the Russian service sector. It seems that the financial reporting of service firms in Russia does not always reflect the real condition of the business. First of all, some operations may be undisclosed, or there may be certain falsifications. Secondly, one firm may comprise several legal entities, and the managers are free to distribute revenues, expenditures, debt and capital between legal entities at will. These factors may make financial reporting biased and, hence, irrelevant to default prediction. Thus, the prediction accuracy may turn out to be low.

In this paper I compared the prediction accuracy of Logit Regression, K-Nearest Neighbors and Random Forest classification algorithms, trained on a set of Russian service firms, as well as on service firms from developed European markets. The algorithms were trained on the financial ratios of defaulted service firms, reported for the year preceding the year of default, and the financial ratios of non-defaulted firms. The firms from developed European markets were used as the control group. It was expected that the accuracy of prediction will be lower for Russian service firms, because of the likely bias in financial reporting, caused by shadow operations and business disaggregation, than for developed European markets' firms, which seem not to have the mentioned features. Hence, the purpose of this study is to estimate the potential default prediction accuracy for Russian service firms if only financial data is used as predictors and to compare it with that for developed European markets' firms. After performing such an analysis, it would be possible to judge whether financial ratios can be used for predicting the default of Russian service firms.

In the next section I provide a review of literature related to default prediction, and subsequently explain why the service sector was selected for this analysis. In the Theoretical framework section I provide a more detailed explanation of why financial ratios do not seem to be reliable default predictors for Russian service firms, and in the Research methods section I describe the data and the algorithms used. Finally, I present and discuss the results of modelling in the "Results" section.

Literature review

Default prediction for firms has existed for over 50 years, starting from the first credit risk model developed by W. Beaver [27]. In an attempt to increase prediction accuracy, it has been evolving in two major domains: methods and explanatory variables.

Firstly, following the development of statistical techniques and econometrics, the researchers started to use more advanced modelling techniques, starting with Edward Altman [1], who implemented Multiple Discriminant Analysis approach, proceeding with James Ohlson [5], who was probably the first to use Logistic Regression to create a default probability assessment model. Logistic Regression (Logit) and a similar algorithm – Probit Regression – were commonly used by the 20th century researchers and are still used nowadays [4; 8; 10; 17], mostly because of their simplicity, given that these are linear algorithms. However, the currently used Machine Learning algorithms seem to be the leading framework in default prediction studies.

There are many different Machine Learning algorithms that are used for default prediction purpose, however, based on the analyzed literature, the most popular are Artificial Neural Networks [14; 28] and Support Vector Machine [18; 25].

One of the contributions of this paper is the implementation of the Random Forest Algorithm as the underlying default prediction technique. This algorithm seems to be underused in default prediction studies, despite its high performance demonstrated by previous researchers [29; 30].

A separate area of research within default prediction is credit rating modelling [31; 32]. The models are based on financial data for corporations and macroeconomic data and is applicable mostly to public firms, because of the significant influence of market capitalization on the credit rating.

The second development vector for default prediction is expanding the set of explanatory variables – going beyond the use of only financial data. This development vector is relatively new, a "novel trend in this field" [21]. According to Altman [23], there was no research in this field for small and medium enterprises at all before 2010.

There are no restrictions on the use of any data available for analyzed firms to predict defaults, and the researchers

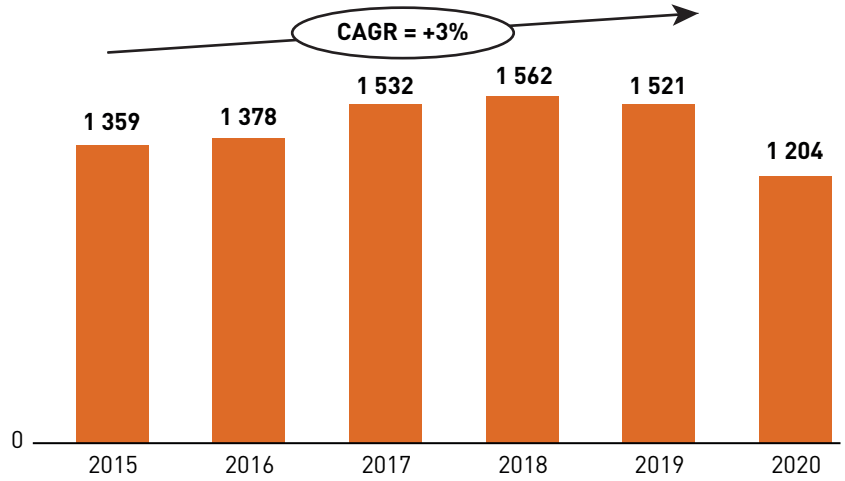
are starting to utilize these available data. The examples of such variables are indicators related to text published in news or disclosures of a firm (e.g. sentiment level or the use of certain words) [25; 26], as well as legal claim-related [21], corporate governance [20], CSR measures [22], and audit report (e.g. sentiment level, number of auditor's comments, etc.) indicators [24].

Based on the analyzed literature use of non-financial data does not seem to replace conventional approach (based on financial data), especially since there are few papers related to Russian firms. This fact can potentially be explained by the high accuracy of default prediction based on financial data. However, as it is demonstrated further in this paper, the approach based on financial data may show poor performance in regard to Russian firms, and in this case the use of non-financial data may prove to be a good solution.

Defaults in the Russian service sector

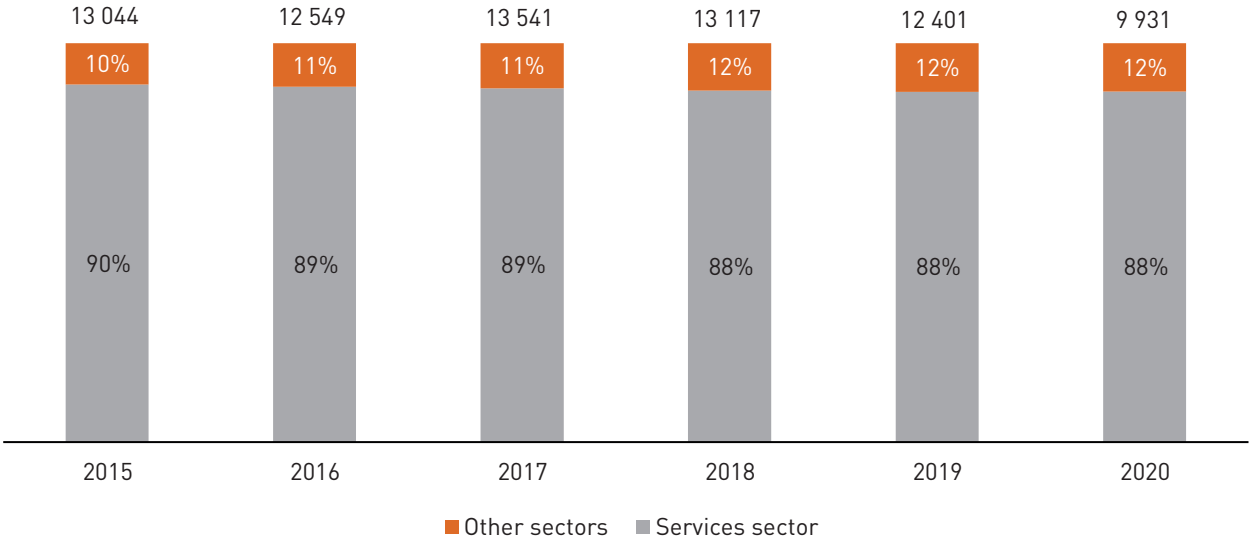
I chose the Russian service sector for this analysis because the need for accurate default prediction is especially relevant in this sector. First of all, in 2015–2020 the overall number of bankruptcies has decreased, while the share of service sector bankruptcies in the overall number of cases increased (see Figures 1 and 2).

Figure 1. Bankruptcies in the Russian service sector, 2015–2020 (number of cases)



Source: Fedresurs. URL: https://fedresurs.ru/news/5343e0f4-bf32-4fef-b293-cc752e65f491 (accessed: 15.06.2021).

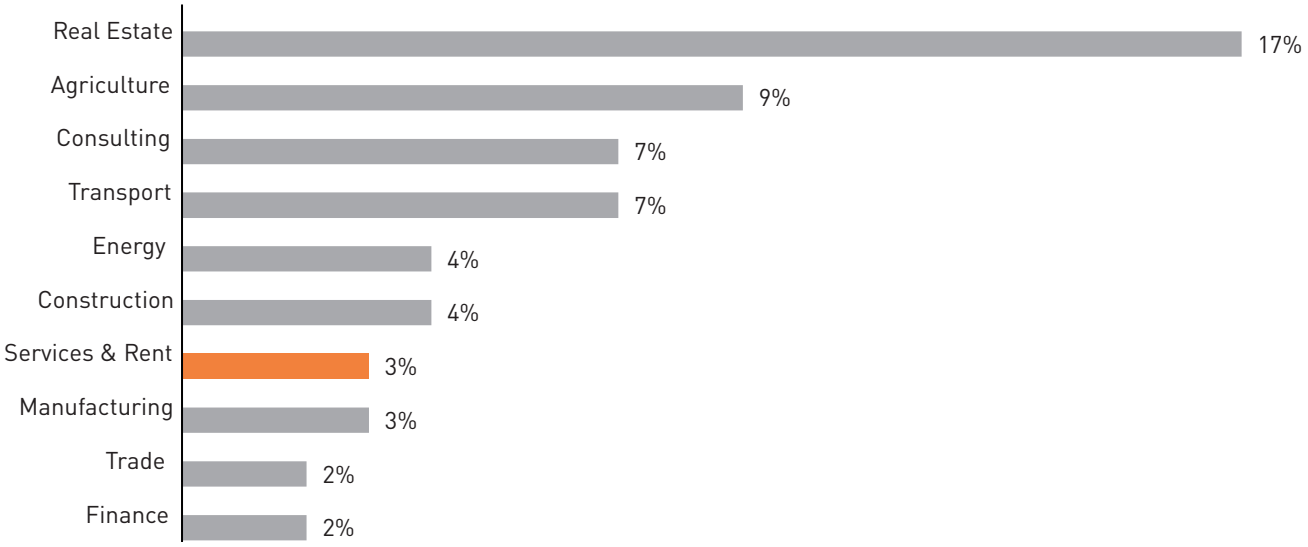
Figure 2. Bankruptcies structure by sector, 2015–2020 (%, number of cases)



Source: Fedresurs. URL: https://fedresurs.ru/news/5343e0f4-bf32-4fef-b293-cc752e65f491 (accessed: 15.06.2021).

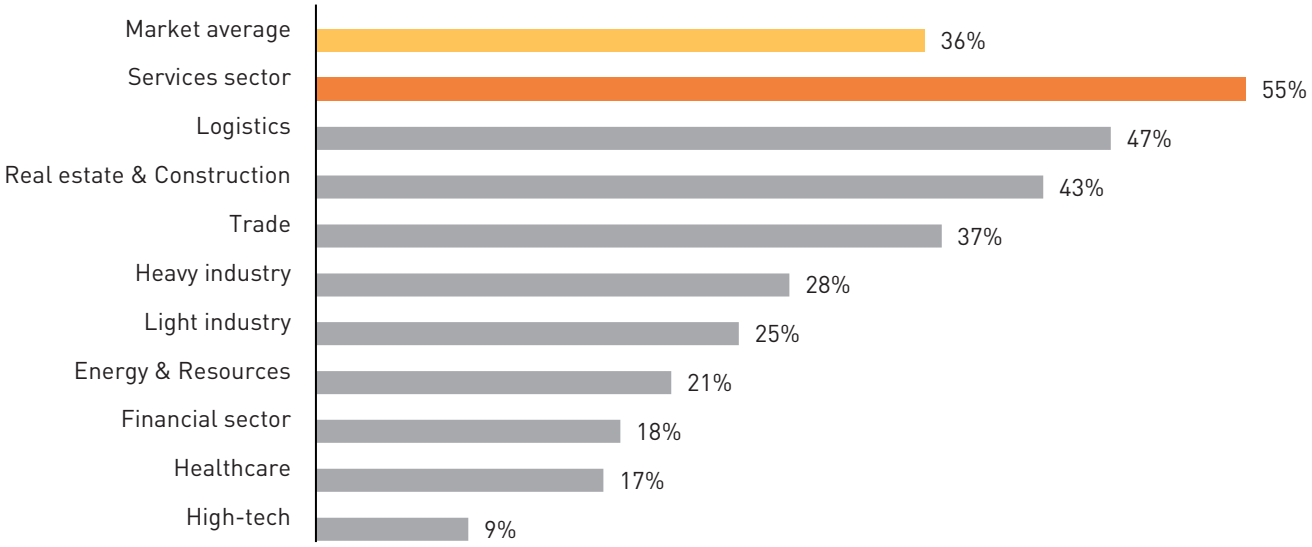
Hereinafter, the year 2020 is not taken into account because of the bankruptcy moratorium in Russia due to COVID-19 pandemic. Secondly, the share of debts paid out to the creditors during default procedures in the service sector is among the lowest across industries. In 2019 this ratio was only 3,4% (less than than the average of 4,7%) (Figure 3). It means that in the case of default the expected amount of debt repayment per 100 RUB borrowed is only 3,4 RUB.

Figure 3. Share of debt paid out in case of default in TOP-10 industries by number of default cases, 2019 (% of total debt)



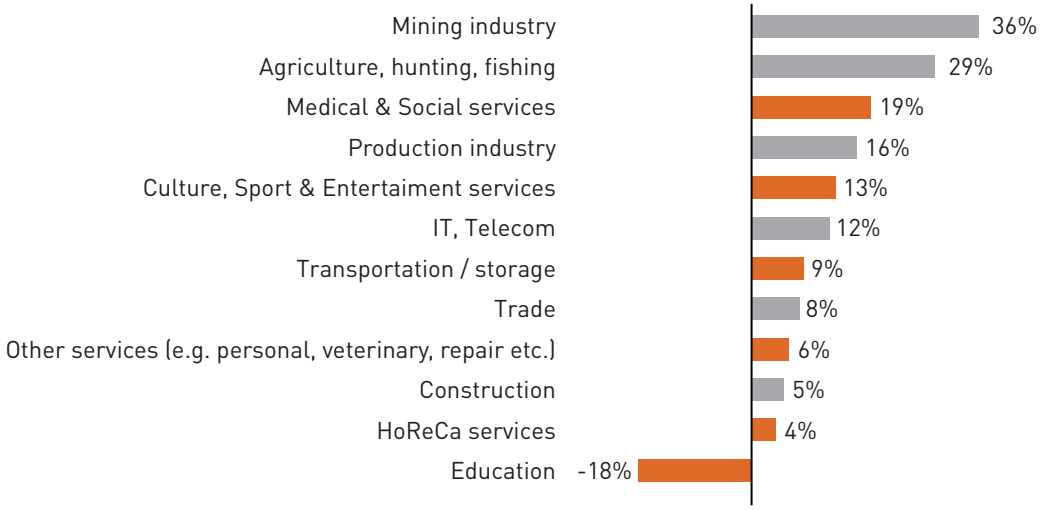
Source: Fedresurs. URL: https://fedresurs.ru/news/5343e0f4-bf32-4fef-b293-cc752e65f491 (accessed: 15.06.2021).

Figure 4. Share of firms with debt by industry, 2020 (%)



Source: Center for Strategic Development.

Figure 5. Return on sales by industry, 2021 (%)



Source: Rosstat. URL: https://www.fedstat.ru/indicator/58261 (accessed: 17.03.2023).

Also, the firms in the service sector tend to have debts more often than those in any other industry. According to the research conducted by Centre for Strategic Development[1], 55% of service firms have debts, while the market average is 36% (Figure 4). This may be an indicator of the higher credit risk of service sector industries compared with other industries.

The increasing number of defaults and the low rate of debt repayment in case of a default are driven by the specificities of the service sector. The sector consists of mostly B2C businesses, which means a high competition level, and therefore low margins. The average profitability of service sector is lower compared to other industries, like production, agriculture, or mining, or even negative (see Figure 5). This statement is less relevant for the medical services sector, but very relevant for such huge markets as HoReCa services and personal services (which include everyday services, i.e., repairs, hairdressing, etc.).

The last but not the least argument to focus on a specific sector of economy, like the service sector, is the gap in the research related to credit risk modelling, which is expressed in the lack of industry focus in such studies as described in [31]. This study aims to contribute to filling the gap for the service sector.

Theoretical framework

The statement that the financial reporting of Russian service firms does not reflect the real condition of the firms is based on two main reasons:

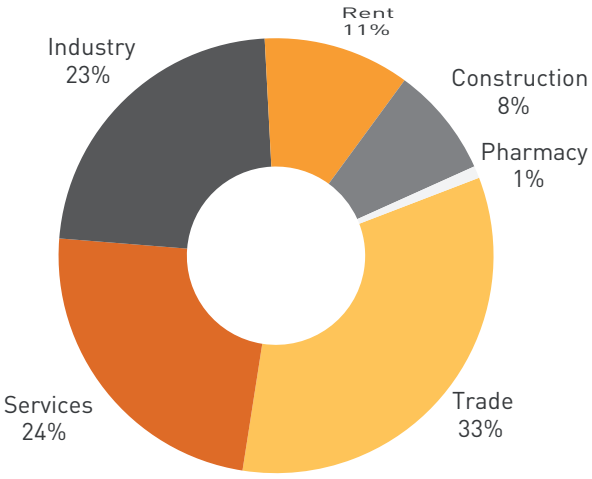
Business disaggregation (artificial separation) make the financial ratios biased

If a firm is divided into several legal entities, it means that it is necessary to obtain the consolidated financial reports in order to judge the condition of the entire business. On the one hand, it is not always possible to get the reports for a group of legal entities, on the other hand, some parts of a group can be presented as sole proprietors or legal entities that use the simplified taxation system and are not required to provide comprehensive reports. That is why one usually has to use data for one legal entity to analyze a firm, and it seems that this data may be biased.

The problem of business disaggregation is highly relevant for the Russian market. Small legal entities have an opportunity to reduce their tax burden using the simplified taxation system. That is why the owners often split their business into several small entities, hence, reducing the tax burden [33]. The relevance of the business disaggregation problem is confirmed by the active prevention measures undertaken by the government. Since 2017, Federal Tax Service and the Investigative Committee of Russia have been actively pursuing a relevant crime detection policy, which includes continuous development and updates to disaggregation criteria [34].

The business disaggregation problem is relevant for every economic sector in Russia, including the service sector. According to a survey conducted by TaxCoach[2], 24% of legal claims on business disaggregation in 2020 were related to service firms (Figure 6).

Figure 6. Legal claims on business disaggregation by industry, 2020 (%)



Source: TaxCoach.

Shadow operations lead to bias in the financial ratios

In the Soviet period, there were no legal private firms in Russia that could provide services for the population. At the same time, government-backed entities did not provide certain everyday services. Thus, the needed services were provided by individuals, including repairs, transport, tutoring, etc. It was an illegal, but the sole way to get the needed services. The prolonged involvement in the shadow economy affected the concept of business culture in the minds of Russian citizens [35]. Moreover, the effect is still apparent.

According to the survey by The Forum for Research on Eastern Europe and Emerging Economies (FREE Network)[3], the volume of the shadow economy in Russia is estimated to be almost 45% of GDP. The two major types

[1] Papchenkova, E. (2020, December 24). *Бизнес-климат России. Итоги 2020 года. Банкротство. (Business climate in Russia. 2020 year summary. Bankruptcy).* URL: https://www.youtube.com/watch?v=cF98nMjWSbs (accessed: 29.05.2021).

[2] TaxCoach. (2021). *Остаться в живых. Гид по обвинениям в искусственном дроблении бизнеса на основе анализа 450 арбитражных дел (Stay alive. A guide for legal claims for artificial business separation, based on 450 legal proceedings)*. URL: https://www.taxcoach.ru/taxbook/droblenie_biznesa (accessed: 01 June 2021).

[3] Putniņš, T., & Sauka, A. (2020). *The Shadow Economy in Russia: New Estimates and Comparisons with Nearby Countries.* URL: https://freepolicybriefs.org/2020/03/16/shadow-economy-russia/ (accessed: 03.06.2021).

of shadow operations are the underreporting of profits and "envelope wages" (according to Tatiana Golikova[4], Deputy Prime Minister of the Russian Federation, about 15 millions of Russian citizens receive wages off the books). According to Russian Longitudinal Monitoring Survey (HSE) 2020[5], 16% of Russian citizens confess being paid off the books, and 51% of them receive their entire salary unofficially.

If a firm is involved in some type of shadow operations, the official financial reporting for a legal entity will be biased: the revenues may be underreported, the costs may be exaggerated etc.

Additional indirect evidence of biased financial reporting by Russian firms is offered by the weak auditing and accounting standards. According to The World Bank Global Competitiveness Index data[6], Russian Federation is in the 100th position out of 137 countries by the strength of auditing and accounting standards (4 out 7 points earned for the question "In your country, how strong are financial auditing and reporting standards? (1 = extremely weak; 7 = extremely strong)).

Thus, these factors lead us to believe that the available financial ratios of Russian services firms may be biased, hence, use of only financial information is not sufficient to assess the credit risk in case of Russian service firms.

Research methods

Data description

It is necessary to specify the industries I consider to be parts of the service sector, because there is no single definition of it. According to *Great Russian Encyclopedia*[7], the service sector includes cultural, educational and domestic services. Russian Federal State Statistics Service[8] identifies postal, telecommunication, housing and utilities, medical and care, tourism, educational, cultural and legal services to be part of public service sector. In this study, I worked with firms from the following industries, which are definitely elements of the service sector:

- Tourism, Accommodation and Passenger Transportation Services;
- Dining & Catering;
- Education;
- Medical & Social Services;
- Culture, Sport & Entertainment Services;
- Other services (personal services, veterinary services, repair services).

The OKVED-2 classification was used to select Russian firms to be included in the analysis, and the NACE Classification was used to choose the European firms. The number of firms by service category is provided in Appendix 1.

I prepared two datasets. The first dataset contains information for Russian service firms, which faced financial failure from 2017 to 2020. The year when the creditor sent out the notice of intent to file an application for default, was used to identify the year of the financial failure. The data was collected from the SPARK-Interfax database,[9] and the dataset consists of 202 failed firms. Each of these firms is paired with a "healthy" one – a firm that has not defaulted. The matching criteria is the value of total firm assets. This matching criteria is commonly used by the researchers [8].

The dependent variable is a dummy variable: 1 stands for defaulted firms, 0 for "healthy" ones. The independent variables are the financial ratios of the firms (calculated for the year preceding the financial failure for defaulted firms).

The most popular financial ratios used by the researchers to create default prediction models, are the following:

- Turnover ratios;
- Profitability ratios;
- Liquidity ratios;
- Assets, equity or debt structure ratios, debt coverage ratios [36].

It turned to be impossible to include debt coverage ratios, because the value of interest payments is not available for the majority of the Russian firms in the dataset. The final list of independent variables used is provided in Table 1.

Table 1. List of independent variables

Turnover ratios	Net assets turnover
	Stock turnover
	Collection period
	Credit period
Profitability ratios	Profit margin
	ROA
Liquidity ratios	Current ratio
	Liquidity ratio
Assets, equity or debt structure ratios	Shareholders' funds / Total Assets

Source: Prepared by the author.

[4] Golikova, T. (2019, June). *Interview with Tatiana Golikova for IZVESTIA*. URL: https://iz.ru/886870/elena-loriia-elena-likhomanova/deistvie-sotckontrakta-ne-dolzhno-ogranichivatsia-mesiatcem-ili-godom (accessed: 03.06.2021).

[5] Russian Longitudinal Monitoring Survey (HSE) 2020. URL: https://www.hse.ru/rlms/spss (accessed: 19.02.2022).

[6] Competitiveness Rankings. (2017). Global Competitiveness Index 2017-2018. URL: http://wef.ch/2wcVUt8 (accessed 21.09.2021).

[7] Big Russian Encyclopedia. URL: https://bigenc.ru/economics/text/3546082 (accessed: 10.06.2021).

[8] Rosstat. URL: https://rosstat.gov.ru (accessed: 10.06.2021).

[9] SPARK Interfax. URL: https://spark-interfax.ru (accessed: 12.06.2021).

Table 2. Descriptive statistics of variables for the two datasets

Variable	Obs	Mean	Std. Dev.	Min	Max
European data, defaults					
Profit margin [%]	145	–11.382	21.112	–97.23	48.447
ROA using Net income [%]	141	–13.629	20.021	–89.8	28.26
Net assets turnover [X]	112	9.984	22.4	.042	183.346
Stock turnover [X]	104	90.301	138.37	2.66	924.534
Collection period [days]	148	51.385	85.619	0	688.013
Credit period [days]	149	61.669	92.285	0	654.728
Current ratio [X]	151	.833	1.069	.005	12.263
Liquidity ratio [X]	148	.755	1.78	.005	12.263
Shareholders' funds / Total Assets [X]	152	–.117	,921	–9.207	.899
European data, non-defaults					
Profit margin [%]	152	4.194	15.772	–83.884	94.162
ROA using Net income [%]	152	5.459	9.087	–20.883	35.16
Net assets turnover [X]	152	5.325	8.147	.06	70.2
Stock turnover [X]	93	122.626	146.115	1.87	845.75
Collection period [days]	152	29.849	36.799	0	213.023
Credit period [days]	152	18.533	19.88	0	108.371
Current ratio [X]	152	2.228	7.108	.014	80.151
Liquidity ratio [X]	152	2.104	7.1	.014	80.151
Shareholders' funds / Total Assets [X]	152	.36	/263	–.607	.987
Russian data, defaults					
Profit margin [%]	201	–937.1	8705.9	–102 815–.1	100
ROA using Net income [%]	202	–462.1	5971	–84837.1	1907.9
Net assets turnover [X]	190	7.961	55.255	–352.55	400.299
Stock turnover [X]	176	372.208	1478.422	0	14753.5
Collection period [days]	199	6600.915	50 664.473	1	579 366
Credit period [days]	195	26 474.502	330 690.63	2.57	4 618 755.6
Current ratio [X]	200	5.654	19.806	.005	180.6
Liquidity ratio [X]	200	4.635	15.212	.003	122.56
Shareholders' funds / Total Assets [X]	200	–5.684	53.369	–750.114	1
Russian data, non-defaults					
Profit margin [%]	202	.051	.302	–2.902	100

Variable	Obs	Mean	Std. Dev.	Min	Max
ROA using Net income [%]	200	–.182	10.05	–132.3	4765.2
Net assets turnover [X]	184	304.016	1554.744	–620.513	16 101/7
Stock turnover [X]	172	1042.513	9181.763	.2	117 718
Collection period [days]	190	182.779	1472.473	1	20 240
Credit period [days]	194	1436.098	18 422.058	.42	256 678.65
Current ratio [X]	202	4.169	11.505	.012	140.883
Liquidity ratio [X]	200	3.361	11.143	.007	140.883
Shareholders' funds / Total Assets [X]	202	–1.259	9.951	–81.2	.993

Source: Prepared by the author.

The second dataset is the control group. It contains the same information, but for service firms from the developed European Union economies (152 defaulted and 152 "healthy" firms). The date of the start of insolvency proceedings was used to identify the year of the financial failure. The data was collected from the Amadeus database[10].

I chose firms from the developed European Union countries as a control group, because the problems of shadow operations and business disaggregation are far less relevant for them. While the shadow market volume in the Emerging & Developing European countries is estimated to be around 27%, the same ratio for the European Union is two times lower (only about 14%)[11]. The countries with the lowest shadow economy ratios are: Austria, Luxembourg, Great Britain, Netherlands, France, Ireland, Island, Germany, Denmark, Sweden, Slovakia, Finland, Spain, Norway[12]. Firms from these countries are used to form the control dataset.

As for busines disaggregation, it seems that there are no statistics for European Union, but one still can state that this problem is less relevant for the European market. Given that business disaggregation is a tool for reducing the tax burden, the attitude of the business community to tax rates can be a proxy for the level of disaggregation. According to the World Bank data[13], 22.6% of Russian firms consider tax rates the biggest obstacle for their business. The same indicator for Austria is only 20.6%, Denmark – 6.4%, Luxembourg – 5.7%, Netherlands – 7.4%, Ireland – 13.6%, Sweden – 13.4%, Slovakia – 17.7%, Finland – 9.5%. There is no data for other European countries on the list, but presumably they are less concerned with the business disaggregation problem, being at a higher "development level." GDP per capita is used as a proxy for the countries' "development level." GDP per capita in the rest of the countries with no data for attitude to taxes is much higher than in Russia,[14]

The variables' descriptive statistics for the two datasets are provided in Table 2. One may notice that Russian financial reporting data has some specificities, e.g. extremely low profitability ratios or extremely high collection and credit periods for defaults. These specificities may be also an indicator of biased financial reporting. A decision was made not to treat the firms with extreme values as outliers, because these extreme values are taken from real financial reporting (the reporting for these firms was checked manually).

Machine Learning algorithms

I used three Machine Learning algorithms to train the data: Logistic Regression, K-Nearest-Neighbors (KNN) and Random Forest. Logistic Regression is a linear classification algorithm that is often used for the purpose of default prediction [5–8]. One of the advantages of Logistic Regression is the ability to interpret the contribution of every independent variable to the prediction. KNN was chosen as probably the most simple machine learning algorithm that is frequently used in studies related to default prediction [37]. The Random Forest classifier was chosen as one of the most powerful algorithms used for default prediction and scoring, as shown in the previous studies [29; 30].

[10] Amadeus Database. (2021). URL: https://amadeus.bvdinfo.com (accessed: 15.07.2021).

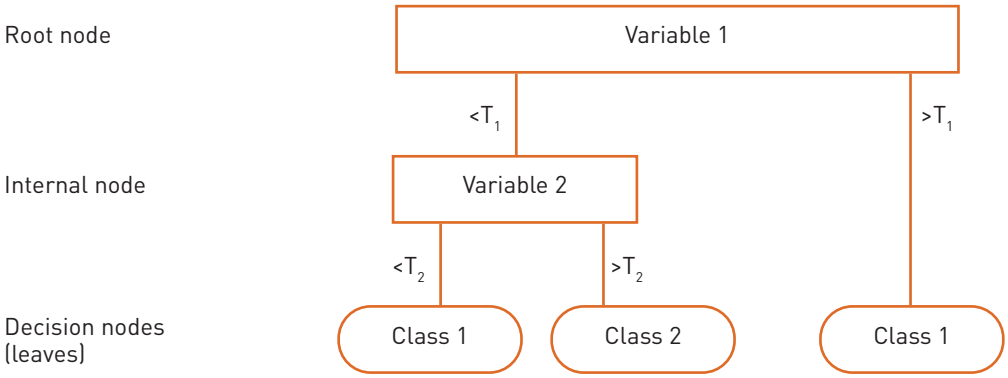
[11] Boumans, D., & Schneider, F. (2019). Ifo World Economic Survey (No. 18; p. 2)]. Leibniz Institute for Economic Research at the University of Munich. URL: https://www.ifo.de/DocDL/WES_4_19_0.pdf (accessed: 08.08.2021).

[12] Kelmanson, B., Kirabaeva, K., Medina, L., Mircheva, B., & Weiss, J. (2019). Explaining the Shadow Economy in Europe: Size, Causes and Policy Options [IMF Working Paper]. International Monetary Fund. URL: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjHiM3o4dvzAhVFpIsKHaXvDLoQFnoECAgQAQ&url=https%3A%2F%2Fwww.imf.org%2F~%2Fmedia%2FFiles%2FPublications%2FWP%2F2019%2Fwpiea2019278-print-pdf&usg=AOvVaw3112V7M9BqTYQptaO-Xh1z (accessed: 10.08.2021).

[13] Enterprise Surveys (The World Bank Data). URL: https://www.enterprisesurveys.org/en/custom-query (accessed: 15.09.2021).

[14] GDP (The World Bank Data). URL: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD (accessed: 15.09.2021).

Figure 7. An example of a simple decision tree (CART)



Source: Prepared by the author.

Logistic Regression is an algorithm that is similar to ordinary linear regression. The difference is that the predicted dependent variable can vary only from 0 to 1, while in the case of ordinary linear regression it can assume any values. For making predictions we use the logistic function (logistic curve):

$$P(x) = \frac{e^{B_0 + B_1 X_1 + \cdots + B_n X_n}}{1 + e^{B_0 + B_1 X_1 + \cdots + B_n X_n}}.$$

$P(x)$ in the case of this study is the estimated probability of default and B_0 – B_n are the linear coefficients for the independent variables (financial ratios). To transform the regression into a classification algorithm, I set the cutoff probability value (50% in this case). The observations are classified into the default group if the estimated probability is higher than 50%.

Logistic Regression is fitted using the maximum likelihood method. The optimal coefficients are chosen in order to maximize the likelihood function:

$$LF = \prod \left(P(x_i)^{y_i} \cdot \left(1 - P(x_i)^{1-y_i}\right)\right), i \in (1; n).$$

which is the product of probabilities, estimated for defaults, and multiplied by the same for non-defaults [38].

The L1 type of regularization is used to limit the number of variables. It means that the sum of absolute values of coefficients is added to the minimized function

The K-Nearest Neighbors classifier is one of the simplest classification algorithms. The classification is based on the classes of several (k) most similar firms from the training set. The observation is classified on the basis of a majority vote. The classification procedure consists of three steps:

- Choosing the number of "neighbors".

The number of "neighbors" should not be very small (may lead to low accuracy) or very high (most of the observations in the test set will be classified as one class, which has more representatives in the training set). I used the square root of the number of observations as k, following the approach recognized by researchers [39].

- Assessing distances between training and test data and identifying the "neighbors".

I use Euclidian distance to choose the nearest "neighbors", calculating it as the following:

$$\sqrt{\sum \left(\begin{matrix} Value\ of\ variable\ i\ for\ the\ observation\ in\ the\ test\ set - \\ Value\ of\ variable\ i\ for\ the\ observation\ in\ the\ train\ set \end{matrix} \right)^2}$$

- Classifying the test observation on a majority vote basis, in other words, assigning a class based on the most popular class among the "neighbors"[15].

Due to the fact that Euclidian distance is used, data needs to be normalized before modelling.

The Random Forest classifier is an ensemble Machine Learning algorithm – an ensemble of Classification and Regression Trees (CART). An illustration of a simple CART is shown in Figure 7.

While the tree is trained, the training data is split into 2 subsamples on every node. The split is made based on a particular variable's value. The Gini index is used to choose the variables (Variable 1, Variable 2 on Figure 7) and the threshold for splitting (T_1, T_2 on Figure 7) – the core idea is to minimize this index. The Gini Index reflects the inverse accuracy of splitting:

$$Gini\ index = \frac{Number\ of\ observations\ in\ L}{Total\ number\ of\ observations} \cdot \left(1 - \sum \left(\frac{Num.\ of\ observations\ of\ class\ i\ in\ L}{Total\ number\ of\ observations\ in\ L} \right)\right) + \frac{Number\ of\ observations\ in\ R}{Total\ number\ of\ observations} \left(1 - \sum \left(\frac{Num.\ of\ observations\ of\ class\ i\ in\ R}{Total\ number\ of\ observations\ in\ R} \right)\right)$$

[15] Laszlo, K. (2008). K Nearest Neighbors algorithm (kNN). Special Course in Computer and Information Science. URL: http://www.lkozma.net/knn2.pdf (accessed: 15.08.2021).

L and R refer to subsample 1 and subsample 2 (left and right), i refers to the class (1 – defaulted, 0 – "healthy) [40].

"Forest" stands for a combination of simple decision trees, "Random" – for the fact that each tree is trained on a randomly chosen subsample from the training sample and the "splitting" variables are chosen randomly. The subsamples are formed using bootstrap. The idea underlying this method is that repeated samples are taken from the initial training sample. For every tree, the variables (Variable 1 and Variable 2 on Figure 7) are chosen from a random list of k variables, taken from the whole list of determinants. Thanks to this, the trees are not similar to each other[16].

It is necessary to limit the number of trees and internal nodes in every tree. It was decided to train 100 trees for each training set and set the maximum number of split layers at 2.

Data preparation and modelling

The datasets contained some missing values. To get rid of them I imputed the data with mean values of the corre-sponding variables. Table 3 shows the fractions of missing values for every variable in two datasets. There are some differences, but it seems that the quality of the collected data is similar for Russian and European firms.

Table 3. Fractions of missing values in the datasets (%)

	Defaulted		Non-defaulted	
	Russian data	European data	Russian data	European data
Net assets turnover	6	26	9	0
Stock turnover	13	32	15	39
Collection period	1	3	6	0
Credit period	3	2	4	0
Profit margin	0	5	0	0
ROA	0	7	1	0
Current ratio	1	1	0	0
Liquidity ratio	1	3	1	0
Shareholders' funds / Total Assets	1	0	0	0

Source: Prepared by the author.

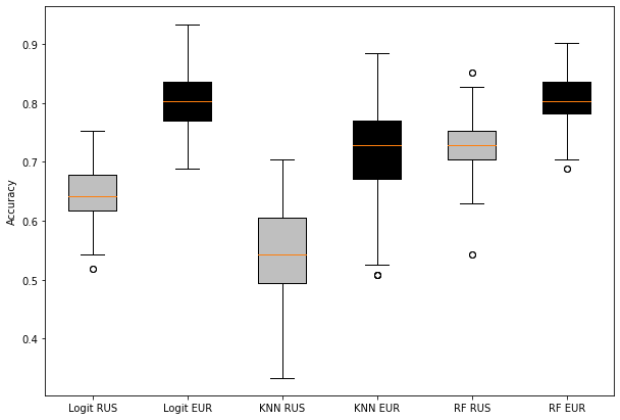
I divided each of the samples (Russian and European firms) into training and test sets. Subsequently, I trained the classification algorithms on the training sets, then applying the trained algorithms to test sets and calculated prediction accuracy. To make sure that the result is not an outlier that occurred because of specific train-test dataset split, I made 100 random train-test splits for every dataset and then trained the algorithms on every training set and calculated the accuracy on every corresponding test set.

The main hypothesis is that the mean accuracy for Russian service firms is going to be lower than for European service firms. This hypothesis was tested using the Mann-Witney test.

Results

The results demonstrate that prediction accuracy is much lower for Russian firms. The results for the three classification algorithms are provided in Figure 8.

Figure 8. Classification results for Logit, KNN and Random Forest algorithms

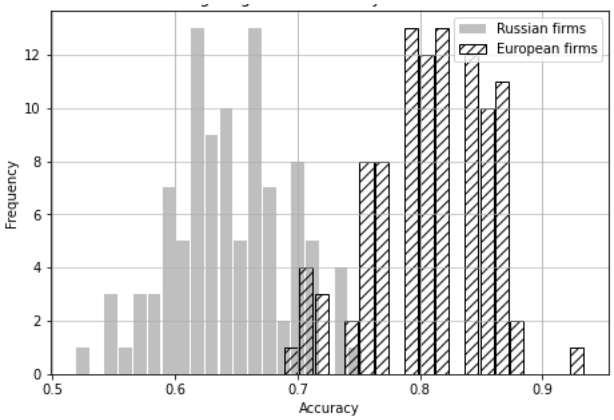


Source: Prepared by the author.

[16] Steorts, R. (2014). Bagging and Random Forests. URL: http://www2.stat.duke.edu/~rcs46/lectures_2015/random-forest/slides_lecture15.pdf (accessed: 15.08.2021).

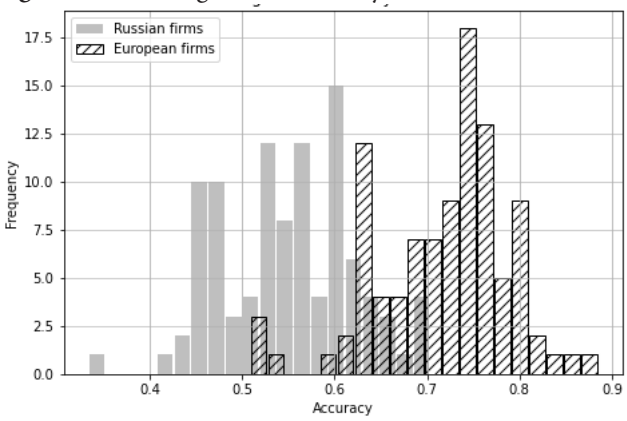
Firstly, I applied Logistic Regression to the datasets. The mean accuracy of classification is 64.4% for Russian service firms and 80.7% for the firms from the European dataset. Figure 9 shows the distribution of Logit algorithm accuracy calculated on the randomly formed test sets for Russian and European service firms. The distribution is visually close to normal in case of both Russian data and European data, but the Shapiro-Wilk normality test result suggests that the accuracies for European firms are not distributed normally (the p-values for Russian and European sets are 0.386 and 0.04 respectively). For instance, the Mann-Witney non-parametric test was used instead of the conventional Student test to test whether the mean accuracies differ. The Mann-Witney test p-value is close to zero ($1.35*10^{-33}$), which means that there is a very low probability of getting such a test statistic if the mean accuracy is the same for Russian and European firms.

Figure 9. Logit algorithm accuracy distribution



Source: Prepared by the author.

Figure 10. KNN algorithm accuracy distribution

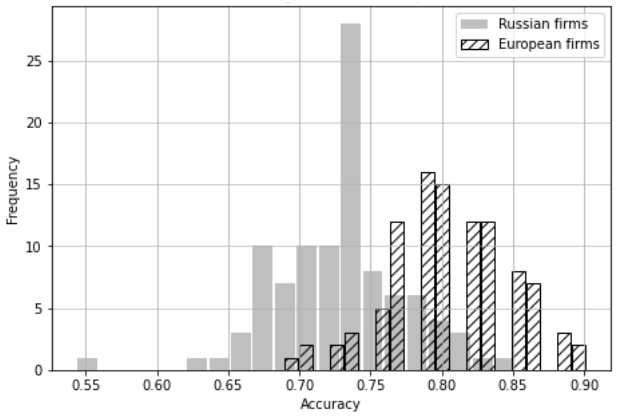


Source: Prepared by the author.

KNN algorithm accuracy is lower in both cases: 54.8% for Russian firms and 71.7% for European firms. Classification accuracy can be considered insufficient for European firms, but it is still significantly higher than the mean accuracy for Russian firms. Figure 10 shows the distribution of KNN algorithm accuracy, calculated on randomly formed test sets for Russian and European service firms. Accuracy distribution is normal in the case of Russian firms, but not in the case of European firms (Shapiro-Wilk test p-values are 0.389 and 0.008 respectively), that is why the Mann-Witney test was used for estimating the significance of the difference in mean accuracies (Figure 10). The p-value of the Mann-Witney test is close to zero ($4.40*10^{-28}$), which means that there is a very low probability of getting such a value if the mean accuracy is the same for Russian and European firms.

The Random Forest algorithm turned to be the most accurate classifier for both Russian and European firms (Figure 11). The mean accuracy of classification is 72.7% and 80.6% for Russian and European service firms, respectively. Figure 11 shows the distribution of Random Forest algorithm accuracy, calculated on randomly formed test sets for Russian and European service firms. The Shapiro-Wilk test results suggest that accuracy distribution is not normal for Russian firms (p-values are 0.019 for Russian firms and 0.18 for European firms), hence I used the Mann-Witney test to assess the significance of the difference in mean accuracies. The p-value of the test is close to zero ($6.14*10^{-23}$), which means that there is a very low probability of getting such a value if the mean accuracy is the same for Russian and European firms.

Figure 11. Random Forest algorithm accuracy distribution



Source: Prepared by the author.

It can be also useful to consider Type I and II errors along with overall accuracy. Table 4 provides the means of Type I and Type II errors for Russian and European datasets according to the algorithm used. The outcomes obtained through overall accuracy analysis are consistent here: both Type I and II errors are bigger in case of Russian service firms, compared with European service firms.

Table 4. Sensitivity, specificity, and Type I & II errors of classification (%)

		Sensitivity	Type I error	Specificity	Type II error
Logit	Russian dataset	72.3	27.7	57.2	42.8
	European dataset	74.9	25.1	86.3	13.7
KNN	Russian dataset	60.5	39.5	51.0	49.0
	European dataset	76.7	23.3	68.7	31.3
Random Forest	Russian dataset	73.4	26.6	72.5	27.5
	European dataset	76.6	23.4	84.6	15.4
Average	Russian dataset	68.7	31.3	60.2	39.8
	European dataset	76.1	23.9	79.9	20.1

Source: Prepared by the author.

Conclusions

Given such results, we can state that default prediction based on financial data is less effective in the case of Russian service firms than in the case of service firms from developed European markets. The accuracy for Russian firms is 55–73%, depending on the algorithm, compared to 72–81% accuracy for the firms from developed European markets. The results for the European dataset in terms of overall accuracy are consistent with the results of previous research [23], while the results for Russian dataset are far behind.

Thus, in case of Russian firms one should expect a higher probability of error while predicting default based on financial indicators. In other words, the results suggest that the financial ratios are worse indicators of future financial failures for Russian firms than for firms from developed markets.

The financial reporting of Russian legal entities does not reflect the real condition of firms due to two possible reasons discussed in this paper: business disaggregation and undisclosed operations. Thus, it may be beneficial to use non-financial factors, which can act as proxies for financial ratios, to improve the accuracy of classification, which can be a starting point for further research related to default prediction in Russia.

Moreover, I believe that the findings of this paper can be generalized in a sense that the conventional approach to default prediction may be inapplicable not only to Russian service firms, but for firms in other developing economies, which are facing the problem of biased financial reporting.

An additional outcome of this study is the comparison of classification algorithms' predictive power. The Random Forest algorithm demonstrates the best performance, supporting the findings of previous research [29; 30]. Despite being a linear classification algorithm, the Logistic Regression classifier can also be used for default prediction (81% accuracy on average for European firms). However, the K-Nearest-Neighbors algorithm seems to be the least accurate (only 72% accuracy on average for European firms and only 55% on average for Russian firms, which means that the predictive power of the algorithm for Russian firms is close to zero).

References

1. Altman E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*. 1968;23(4):589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x
2. Matenda F.R., Sibanda M., Chikodza E., Gumbo V. Corporate default risk modeling under distressed economic and financial conditions in a developing economy. *Journal of Credit Risk*. 2021;17(1):89-115. https://doi.org/10.21314/JCR.2020.267
3. Fedorova E., Musienko S., Fedorov F. Analysis of the external factors influence on the forecasting of bankruptcy of Russian companies. *Vestnik Sankt-Peterburgskogo universiteta. Ekonomika = St Petersburg University Journal of Economic Studies*. 2020;36(1):117-133. (In Russ.). https://doi.org/10.21638/spbu05.2020.106
4. Grigoriev A., Tarasov K. Corporate bankruptcy prediction using the principal components method. *Journal of Corporate Finance Research*. 2019;13(4):20-38. https://doi.org/10.17323/j.jcfr.2073-0438.13.4.2019.20-38
5. Ohlson J.A. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*. 1980;18(1):109-131. https://doi.org/10.2307/2490395
6. Hunter J., Isachenkova N. Failure risk: A comparative study of UK and Russian firms. *Journal of Policy Modeling*. 2001;23(5):511-521. https://doi.org/10.1016/S0161-8938(01)00064-3
7. Gruszczyński M. Financial distress of companies in Poland. *International Advances in Economic Research*. 2004;10(4):249-256. https://doi.org/10.1007/BF02295137
8. Sirirattanaphonkun W., Pattarathammas S. Default prediction for small-medium enterprises in emerging market: Evidence from Thailand. *Seoul Journal of*

Business. 2012;18(2):25-54. https://doi.org/10.35152/snusjb.2012.18.2.002

9. Ahmadpour Kasgari A., Divsalar M., Javid M.R., Ebrahimian S.J. Prediction of bankruptcy Iranian corporations through artificial neural network and Probit-based analyses. *Neural Computing and Applications*. 2013;23(3-4):927-936. https://doi.org/10.1007/s00521-012-1017-z
10. Kovacova M., Kliestik T. Logit and Probit application for the prediction of bankruptcy in Slovak companies. *Equilibrium*. 2017;12(4):775-791. https://doi.org/10.24136/eq.v12i4.40
11. Odom M., Sharda R. A neural network model for bankruptcy prediction. In: 1990 IJCNN International joint conference on neural networks (San Diego, CA, 17-21 June 1990). Piscataway, NJ: IEEE; 163-168. https://doi.org/10.1109/IJCNN.1990.137710
12. Coats P.K., Fant L.F. Recognizing financial distress patterns using a neural network tool. *Financial Management*. 1993;22(3):142-155. https://doi.org/10.2307/3665934
13. Altman E.I., Marco G., Varetto F. Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks (the Italian experience). *Journal of Banking & Finance*. 1994;18(3):505-529. https://doi.org/10.1016/0378-4266(94)90007-8
14. Zhang G., Hu M.Y., Patuwo B.E., Indro D.C. Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*. 1999;116(1):16-32. https://doi.org/10.1016/S0377-2217(98)00051-4
15. Kumar P.R., Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. *European Journal of Operational Research*. 2007;180(1):1-28. https://doi.org/10.1016/j.ejor.2006.08.043
16. Cao Y., Liu X., Zhai .J, Hua S. A two-stage Bayesian network model for corporate bankruptcy prediction. *International Journal of Finance & Economics*. 2022;27(1):455-472. https://doi.org/10.1002/ijfe.2162
17. Mselmi N., Lahiani A., Hamza T. Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis*. 2017;50:67-80. https://doi.org/10.1016/j.irfa.2017.02.004
18. Xie C., Luo C., Yu X. Financial distress prediction based on SVM and MDA methods: The case of Chinese listed companies. *Quality & Quantity*. 2011;45(3):671-686. https://doi.org/10.1007/s11135-010-9376-y
19. Shumway T. Forecasting bankruptcy more accurately: A simple hazard model. *The Journal of Business*. 2001;74(1):101-124. https://doi.org/10.1086/209665
20. Fernando J.M.R., Li L., Hou G. Financial versus non-financial information for default prediction: Evidence from Sri Lanka and the USA. *Emerging Markets Finance and Trade*. 2020;56(3):673-692. https://doi.org/10.1080/1540496X.2018.1545644
21. Blanco-Oliver A., Irimia-Diéguez A., Oliver-Alfonso M., Wilson N. Improving bankruptcy prediction in micro-entities by using nonlinear effects and non-financial variables. *Finance a úvěr – Czech Journal of Economics and Finance*. 2015;65(2):144-166. URL: http://journal.fsv.cuni.cz/storage/1321_blanco_oliver.pdf
22. Boubaker S., Cellier A., Manita R., Saeed A. Does corporate social responsibility reduce financial distress risk? *Economic Modelling*. 2020;91:835-851. https://doi.org/10.1016/j.econmod.2020.05.012
23. Altman E.I., Sabato G., Wilson N. The value of non-financial information in SME risk management. *Journal of Credit Risk*. 2010;6(2):95-127. https://doi.org/10.21314/JCR.2010.110
24. Muñoz-Izquierdo N., Laitinen E.K., Camacho-Miñano M del-Mar, Pascual-Ezama D. Does audit report information improve financial distress prediction over Altman's traditional Z-Score model? *Journal of International Financial Management & Accounting*. 2020;31(1):65-97. https://doi.org/10.1111/jifm.12110
25. Makeeva E., Sinilshchikova M. News sentiment in bankruptcy prediction models: Evidence from Russian retail companies. *Journal of Corporate Finance Research*. 2020;14(4):7-18. https://doi.org/10.17323/j.jcfr.2073-0438.14.4.2020.7-18
26. Feng M., Shaonan T., Chihoon L., Ling M. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*. 2019;274(2):743-758. https://doi.org/10.1016/j.ejor.2018.10.024
27. Beaver W.H. Financial ratios as predictors of failure. *Journal of Accounting Research*. 1966;4:71-111. https://doi.org/10.2307/2490171
28. Zhu Y., Xie C., Wang G.-J., Yan X.-G. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*. 2017;28(1):41-50. https://doi.org/10.1007/s00521-016-2304-x
29. Barboza F., Kimura H., Altman E. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*. 2017;83:405-417. https://doi.org/10.1016/j.eswa.2017.04.006
30. Brown I., Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*. 2012;39(3):3446-3453. https://doi.org/10.1016/j.eswa.2011.09.033

31. Karminsky A. Corporate rating models for emerging markets. *Korporativnye finansy = Journal of Corporate Finance Research*. 2011;5(3):19-29. (In Russ.). https://doi.org/10.17323/j.jcfr.2073-0438.5.3.2011.19-29
32. Grishunin S., Egorova A. Comparative analysis of the predictive power of machine learning models for forecasting the credit ratings of machine-building companies. *Journal of Corporate Finance Research*. 2022;16(1):99-112. https: //doi.org/10.17323/j.jcfr.2073-0438.16.1.2022.99-112
33. Kachalin D. Analysis of Russian models of splitting (reorganization) of business that ensure compliance of its scale with special taxation regime. *Finansovaya analitika: problemy i resheniya = Financial Analytics: Science and Experience*. 2011;(5):52-63. (In Russ.).
34. Donich S.R. Novelties in the tax administration: The concept of splitting a business. *Vestnik Sibirskogo gosudarstvennogo universiteta putei soobshcheniya: Gumanitarnye issledovaniya = The Siberian Transport University Bulletin: Humanitarian Research*. 2021;(1):39-44. (In Russ.).
35. Williams C.C., Nadin S., Newton S., Rodgers P., Windebank J. Explaining off-the-books entrepreneurship: A critical evaluation of competing perspectives. *International Entrepreneurship and Management Journal*. 2013;9(3):447-463. https://doi.org/10.1007/s11365-011-0185-0
36. Jaki A., Ćwięk W. Bankruptcy prediction models based on value measures. *Journal of Risk and Financial Management*. 2021;14(1):6. https://doi.org/10.3390/jrfm14010006
37. Jandaghi G., Saranj A., Rajaei R., Ghasemi A., Tehrani R. Identification of the most critical factors in bankruptcy prediction and credit classification of companies. *Iranian Journal of Management Studies*. 2021;14(4):817-934. https://doi.org/10.22059/IJMS.2021.285398.673712
38. Hosmer D.W. Jr., Lemeshow S., Sturdivant R.X. Introduction to the logistic regression model. In: Applied logistic regression. Hoboken, NJ: John Wiley & Sons, Inc.; 2013:1-33. (Wiley Series in Probability and Statistics). https://doi.org/10.1002/9781118548387.ch1
39. Hassanat A.B., Abbadi M.A., Altarawneh G.A., Alhasanat A.A. Solving the problem of the K parameter in the KNN classifier using an ensemble learning approach. *International Journal of Computer Science and Information Security*. 2014;12(8):33-39. https://doi.org/10.48550/arXiv.1409.0919
40. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5-32. https://doi.org/10.1023/A:1010933404324

Appendix 1

Number of firms in the datasets by services category

Services category	European data		Russian data	
	Number of firms	Fraction of firms, %	Number of firms	Fraction of firms, %
Dining & Catering	91	30	172	43
Other services	52	17	45	11
Medical & Social services	49	16	58	14
Tourism, accomodation and passenger transportation services	72	24	53	13
Culture, Sport & Entertainment services	21	7	66	16
Education	19	6	10	2

The article **was submitted** 25.12.2022; **approved after reviewing** 23.01.2023; **accepted for publication** 20.02.2023.