V. V. Afanasev
Lecturer National research university «Higher school of economics», Saint-Petersburg

DEFAULT PREDICTION FOR AUTO REPAIR SERVICES FIRMS IN RUSSIA USING NON-FINANCIAL DATA

1. Introduction

Conventional default risk assessment models, which has been developed starting from 1960-s (e.g. Altman, 1968; Beaver, 1966; Ohlson, 1980; Zmijewski, 1984), assume that a firm's deteriorating financial condition indicates potential default in the future. Financial ratios have been proved to explain future defaults with high accuracy.

However, if the reporting of a firm is biased due to some kind of manipulation or fraud, biased financial data can lead to poor default prediction. Some researchers are now using non-financial data, to improve the accuracy of default prediction (e.g. Altman, Sabato, Wilson, 2010; Bhimani, Gulamhussen, Lopes, 2013; Lugovskaya, 2010; Xie, Luo, Yu, 2011). The researchers find evidence that using non-financial data can improve the quality of default prediction.

The use of non-financial data for default prediction seems to be highly applicable to the emerging markets. As it is shown in (Afanasev, 2023), the accuracy of the models, based on financial ratios, is significantly higher for developed countries from European Union, than for such an emerging market, as Russia. However, up to now there were a few studies on the use of non-financial data for default prediction in emerging markets.

This study aims to test, whether the use of non-financial data can improve the quality of prediction for auto repair firms in such an emerging market, as Russia. I raise two research questions:

RQ1: To what extent can the accuracy of default prediction be improved, if the model based on financial ratios is enriched with non-financial data?

RQ2: What non-financial factors are the most powerful in terms of default prediction?

The study reveals that financial ratios are insufficient for predicting default for private auto repair firms, and non-financial variables, mostly sum and number of legal claims, significantly improve prediction accuracy. The findings of this study are significant for credit institutions and businesses (mostly counterparties of auto repair firms), as they indicate potential area for improvement of the existing approaches of credit risk assessment, and, furthermore, may be an indicator of the poor quality of financial reporting of such firms.

2. Literature review

2.1. Approaches to the definition of default

It is necessary to discuss the meaning, which is put into the term 'default' in this study, because this meaning differs across different research studies. Firms do not become bankrupt immediately, the process typically starts with non-payment and culminates in official bankruptcy, which can take a considerable amount of time. Also, after the official bankruptcy, a long time may pass before the closure of the firm. In the same time the closure may take place even without any official bankruptcy procedure, or be

DOI 10.31085/1814-4802-2024-20-3-144-71-88

caused by reasons, not related to insolvency, that is why there is a need to distinguish between failure and closure. Consequently, the definition of default lacks clarity and it is necessary to give specific clarification in this study.

A dive into the existing literature reveals that one of the approaches is to treat the firm as a defaulting one if its financial ratios satisfy some condition (e. g. EBITDA less than interest obligations) (Andrade, Kaplan, 1998). But other studies use the start of insolvency proceedings, liquidation, administration or receivership as the indicator of the default (Altman, Sabato, Wilson, 2010; Muñoz Izquierdo et al., 2020). For public firms it can be the delisting date (Mai et al., 2019).

The approach used in this study is similar and implies using the legal claim for bankruptcy application as the indicator of the default, and the date of this application as the default date.

2.2. Previous research in default prediction based on financial data

The development of default risk assessment models began in 1966 with William Beaver's work (Beaver, 1966), followed by research papers by Edward Altman (Altman, 1968), James Ohlson (Ohlson, 1980), Mark Zmijewski (Zmijewski, 1984), who are regarded as pioneers in this field. More recent studies have also focused on predicting defaults using financial data, including very recent papers (Bazzana, Bee, Hussin Adam Khatir, 2024; Cao et al., 2020; Zhao, Lin, 2023).

The underlying assumption in all the mentioned studies is that a firm's deteriorating financial condition today serves as an indication of potential insolvency in the future. The researchers usually use turnover, profitability, liquidity, assets, equity and debt structure ratios and debt coverage ratios (Jaki, Ćwięk, 2021).

However, a firm's deteriorating financial condition can indicate insolvency in the future only if the financial data is taken from the true reporting of the firm, but the reporting can be biased. The cases of fraud in financial reporting seem to be frequent (Hogan et al., 2008). In case of deteriorating financial condition firms can try to delay or hide negative data in the financial reporting (Schwartz, Menon, 1985). Managers can manipulate the financial reporting due to, among other things, personal need for money (Bagherpour Velashani, Jadaan Sallal, Salehi, 2022).

The biased financial data can lead to poor quality of default prediction. As it is shown in (Afanasev, 2023), the accuracy of default prediction based on financial data is significantly lower for firms from an emerging economy, which face problems with reporting quality, than for firms from leading developed European economies. One potential way to increase the accuracy of default prediction is to use non-financial data along with financial ratios, which is a 'novel trend' in default prediction (Blanco et al., 2015).

2.3. Previous research in default prediction based on non-financial data

In recent years, many researchers in default prediction field have begun including non-financial data in their models to improve the accuracy of the models. In default prediction there are a wide selection of variables that are not related to financial ratios, and such variables may be chosen in a way that is only limited by common sense based on the circumstances. A number of variables could be taken into account, such as legal claims data (Shumway, 2001), corporate governance indicators and CSR indicators (Boubaker et al., 2020; Xie, Luo, Yu, 2011), text sentiment in public relations (Mai et al., 2019), as well as measures based on audit reports (if the firm is audited or not, number of comments in the report, text sentiment, etc.) (Altman, Sabato, Wilson, 2010; Blanco et al., 2015).

Even though it is still not a common approach to predict defaults using non-financial data, the researchers have received promising results in terms of prediction accuracy increase when adding non-financial variables in the model. For instance, the increase of the area under the ROC curve by 7–21% (Altman, Sabato, Wilson, 2010; Bhimani, Gulamhussen, Lopes, 2013; Wilson, Ochotnický, Káčer, 2016) or the increase of the overall accuracy by 5–13 percentage points (Laitinen, 2011; Lin, Liang, Chu, 2010; Lugovskaya, 2010; Xie, Luo, Yu, 2011).

It seems that using non-financial data is beneficial for the quality of default prediction. However, the accuracy can also depend on the modelling technique. The next section of the literature review is devoted to summarizing the existing methodologies of modelling defaults.

2.4. Techniques used for modelling defaults

There are two groups of techniques used for default prediction: the so called 'statistical' and 'intelligent' (Kumar, Ravi, 2007). The first group includes such techniques as MDA (Altman, 1968), logistic or probit regression (Gruszczyński, 2004; Laitinen, 2011; Ohlson, 1980; Zhao, Lin, 2023), which are the most commonly used, and other statistical techniques, as linear discriminant analysis, quadratic discriminant analysis etc.

The second group includes machine learning algorithms, such as neural networks, evolutionary algorithms, decision trees based algorithms etc. (Kumar, Ravi, 2007). According to the existing literature, defaults can potentially be predicted more accurately by using machine learning algorithms, as shown in (Ahmadpour Kasgari et al., 2013; Barboza, Kimura, Altman, 2017). However, such methods as logistic or probit regressions have the advantage of the researcher being able to analyze the effect of each variable on the classification result, in contrast to the opaque machine learning algorithms, often realized in a form of a 'black box', e. g. Artificial Neural Networks.

Fortunately, there are some machine learning algorithms, which allow to analyze the contribution of every variable in the classification process. One of such algorithms is Random Forest, described in (Breiman, 2001). It combines high predictive power for default prediction tasks (Barboza, Kimura, Altman, 2017; Brown, Mues, 2012) with the ability to calculate the 'importances' for every of the explanatory variables used in the model, as it is shown, for instance, in (Uddin et al., 2022) or (Bazzana, Bee, Hussin Adam Khatir, 2024). The specifics related to Random Forest algorithm are given in the 'Methodology' section.

2.5. Approaches to model training

A special attention is paid to correct and transparent calculation of the explanatory variables. The existing studies often take the most recent reports, available on the default date. For instance, in (Zmijewski, 1984) the financial data for defaulting firms were 'collected from the last SEC 10K report filed before the date a firm petitioned for bankruptcy'. Besides, there are studies, which give limited explanations of what period is used to calculate the values of the predictors. Usually, the researchers state taking data for 'X years before the default', e. g. 'first' or 'second pre-bankruptcy year' in (Lukason, Hoffman, 2014) or 'one year and two years before the judgment date of the event declaration' in (Mselmi, Lahiani, Hamza, 2017). Furthermore, there are many papers with no explanation given on the period for data collection.

This study provides two approaches: 1) taking the data available on the date one year before the default and 2) taking the most recent available data on the date of default. In 'Data description' section I explain, which periods are used to calculate the values for the predictors, in details.

3. Methodology

3.1. Data description

The dataset I used for modelling consists of firms under NACE Class 45.2 — Maintenance and repair of motor vehicles. The dataset consists of 2 groups of firms: those, which faced a default, and those, which did not face a default (hereinafter — healthy peers). I treat a firm as a defaulted one, if there was a lawsuit filed for bankruptcy. The date of default is assumed to be the same as the date of this lawsuit filing. The initial default dataset consisted of 305 firms, which faced a default in the period from 2018 to 2023.

The dependent variable is a binary one: 1 if the firm is in the default group, 0 — otherwise. The list of independent variables consists of two groups: financial and non-financial. The financial data was collected from SPARK INTERFAX database for private firms. The set of financial ratios was chosen to cover 4 groups of financial ratios: profitability, liquidity, solvency, turnover (see Table 1).

Table 1

Financial independent variables

Group of ratios	Variable name	Description
Profitability	GP2A	Gross profit divided by total assets
	ROS	Return on sales
	ROE	Return on equity
Liquidity	CLR	Current liquidity ratio
Solvency	D2E	Gearing ratio
Turnover	WC2A	Working capital divided by total assets
	Rec2A	Receivables divided by total assets
	A_turn	Total assets turnover
	Rec_turn_days	Accounts receivable turnover in days
	Pay_turn_days	Accounts payable turnover in days

Table 2

Non-financial independent variables

Variable name	Description
N_inspections	The number of inspections, faced by the firms during 1 year before the theoretical forecast date
N_violations	The number of violations, found during the inspections, faced by the firms during 1 year before the theoretical forecast date
N_tender_wins	Number of won tenders (government procurement) during 1 year before the theoretical forecast date
N_management_or_shareholders_changes	Number of changes in management or shareholders during 1 year before the theoretical forecast date
N_location_changes	Number of changes of location during 1 year before the theoretical forecast date
N_claims	Number of legal claims filed against the firm during 1 year before the theoretical forecast date
Sum_claims	Sum of legal claims filed against the firm (in rubles) during 1 year before the theoretical forecast date

The set of non-financial variables was chosen based on the availability of data related to private firms on the one hand, and the possibility to collect the data for a big set of firms on the other hand. The main source for non-financial data is Vbankcenter.ru website — a marketplace for tender procurement with an option to get additional non-financial data about any legal entity. The number and the value of legal claims were obtained from SPARK INTERFAX database for private firms. The set of non-financial variables is given in Table 2.

To form the 'healthy' dataset it was decided to pair every defaulted firm with a random healthy peer, using the value of total assets as the matching criterion. Given that the total number of legal entities, registered in the industry, is more than 21 000, it was decided to lower the sampling error risk by creating 10 groups of healthy peers and fitting the model 10 times to check if the results are consistent across different samples. I randomly selected 10 peer observations for each default observation so that both were in the same percentile of total assets in the corresponding period. There can be repetitions (e. g. the same firm may be a peer for a firm, which faced a default in 2018 and a peer for a firm which faced a default in 2019, in this case the data for the healthy peer is taken for different years), because the peer sample is constructed randomly using all potential peers.

I used two approaches to choose the time period, for which the values of independent variables should be calculated. The first approach was to model the default prediction one year before the date of default. Hereinafter I call this approach the 'Year-before-approach'. A model build under this approach can be used to predict defaults in one year (the theoretical forecast date is exactly one year before the default). The second approach was to take the most recent available financial reporting on the default date to calculate the financial variables, which means that the theoretical forecast date is the date, on which this financial report is becoming available. The annual financial reporting in Russia should be submitted before 31 March, that is why I treated 01 April as the theoretical forecast date (e. g. for a firm, which faced a default on 01 June 2022, the theoretical forecast day should be 01 April 2022, but for a firm, which faced a default in March 2022, the theoretical forecast date should be 01 April 2021). I call the second approach the 'Last-available-financial-report-approach'. The default sample is the same under both approaches, but the healthy peers samples are formed independently and, thus, are different. The principles of independent variables calculation under the two approaches are given in Table 3.

Table 3

Principles of independent variables calculation

Variables type:	**Modelling approach:**	**'Year-before-approach'**	**'Last-available-financial-report-approach'**
Financial	*Default firms*	The financial ratios are calculated based on the financial reporting available one year before the default	The financial ratios are calculated based on the financial reporting available on the default date
	Healthy peers	The financial ratios for healthy peers are calculated for the same period as the ratios for their default peers	The financial ratios for healthy peers are calculated for the same period as the ratios for their default peers
Non-financial	*Default firms*	The non-financial variables are calculated for 12 months ending on the theoretical forecast date (one year before the default)	The non-financial variables are calculated for 12 months ending on the theoretical forecast date (most recent 31 March — the date of financial reporting submission)
	Healthy peers	The non-financial variables are calculated for 12 month ending on 01 July of the year, for which the financial ratios for this firm are calculated	The non-financial variables are calculated for 12 months ending on the theoretical forecast date (most recent 31 March — the date of financial reporting submission)

Some of the financial ratios were missing for several firms. It was decided to exclude observations with more than 5 financial ratios missing (half of all financial ratios). Other missing values were imputed with the mean values of the corresponding variable. In the end, the total number of observations appeared to be 2712 under the 'Year-before-approach' and 2240 under the 'Last-available-financial-report-approach' (see Table 4).

Table 4

Number of observations in the samples

Sample	'Year-before-approach'	'Last-available-financial-report-approach'
Default sample	256	221
Healthy peers sample 1	246	202
Healthy peers sample 2	246	202
Healthy peers sample 3	245	201
Healthy peers sample 4	246	202
Healthy peers sample 5	246	202
Healthy peers sample 6	244	202
Healthy peers sample 7	246	202
Healthy peers sample 8	246	202
Healthy peers sample 9	245	202
Healthy peers sample 10	246	202
Total	2712	2240

The default sample consists of firms, which faced the default in different time periods. The distribution of the final default sample by the year of the default is given in Table 5.

Table 5

The distribution of the final default sample by the year of the default

	'Year-before-approach'		'Last-available-financial-report-approach'	
Year	Number of default firms	% of total	Number of default firms	% of total
2018	22	9%	22	10%
2019	39	16%	44	20%
2020	40	14%	35	16%
2021	114	45%	84	38%
2022	36	14%	31	14%
2023	5	2%	5	2%
Total	256	100%	221	100%

3.2. Classification algorithm

I chose Random Forest Classification algorithm, described in (Breiman, 2001) for the purpose of modelling because of two reasons. Firstly, this algorithm has shown high predictive accuracy in previous studies by several researchers (Barboza, Kimura, Altman, 2017; Brown, Mues, 2012). Secondly, it is one of the few algorithms, which has an option to assess the contribution of the independent variables in the explanation of the dependent variable, which is a necessary option if one aims to understand, whether non-financial factors improve the accuracy of default prediction.

The Random Forest classification algorithm is based on simple classification trees but uses an ensemble of such trees (a 'forest'). Every classification tree assigns a class for an observation, and the most frequent assigned class is then treated as the output of the 'forest'. In the case of this research, I use a 'forest' of 555 trees.

Bootstrapping method is used to train the trees. The training of each tree is made on randomly picked subsets (with replacement) from the training data (this makes the 'forest' to be 'random'). The size of such subsets in case of this study are the same as the number of observations in the training set, but every tree is trained on a slightly different subset, because of the replacement option (one observation can be included in a subset several times).

The core aim of a classification tree while training is to split the training data so that every leaf of the tree contains only one class of observations, or the share of this class is as high as possible. There can be several split nodes in a tree, and every node splits the data into two subsamples (denoted as L and R — left and right), minimizing the Gini index:

$$
\begin{aligned}
Gini\ index = {} & \frac{\#\,of\ observations\ in\ L}{Total\ \#\ of\ observations\ to\ split} \\
& \cdot \left(1 - \sum \left(\frac{\#\,of\ observations\ of\ class\ i\ in\ L}{Total\ \#\ of\ observations\ in\ L}\right)^2\right) \\
& + \frac{\#\,of\ observations\ in\ R}{Total\ \#\ of\ observations\ to\ split} \\
& \cdot \left(1 - \sum \left(\frac{\#\,of\ observations\ of\ class\ i\ in\ R}{Total\ \#\ of\ observations\ in\ R}\right)^2\right)
\end{aligned}
$$

Table 6

The hyperparameters used for Random Forest

Hyperparameter	*Value*
Number of trees	555
Maximum number of tree split layers	5

4. Empirical results and discussion

The results show that the accuracy of the default prediction for Russian auto repair firms is low if one uses only financial ratios to predict the default. This result holds for both the 'Year-before-approach' and the 'Last-available-financial-report-approach'. The mean accuracy of classification on 10 test sets is 62% and 64% respectively. The sensitivity is much higher than specificity (70% and 68% sensitivity depending on the approach and 54% and 60% specificity).

However, when non-financial data is added to the model, the mean accuracy of default prediction on 10 test sets goes up by around 10 percentage points (72% for the 'Year-before-approach' and 73% for the 'Last-available-financial-report-approach'). Also, the sensitivity and specificity are both around 70–73%, which indicates better 'balance' of the classification algorithm. The most powerful non-financial variables in terms of classification are the sum and the number of legal claims against the firm.

Thus, the results show that non-financial factors can significantly improve the quality of default prediction. The rest of this section serves to give a more detailed description of the results.

4.1. 'Year-before-approach', model with financial data only

Firstly, I trained Random Forest algorithm on 10 training samples, formed under 'Year-before-approach'. The training was implemented using financial variables only. The mean accuracy of default prediction on the training samples is 85.6% (see Figure 1), the mean accuracy on the test samples is 62.2% (see Figure 2). The average area under ROC curve is 0.94 for training data and 0.68 for test data.

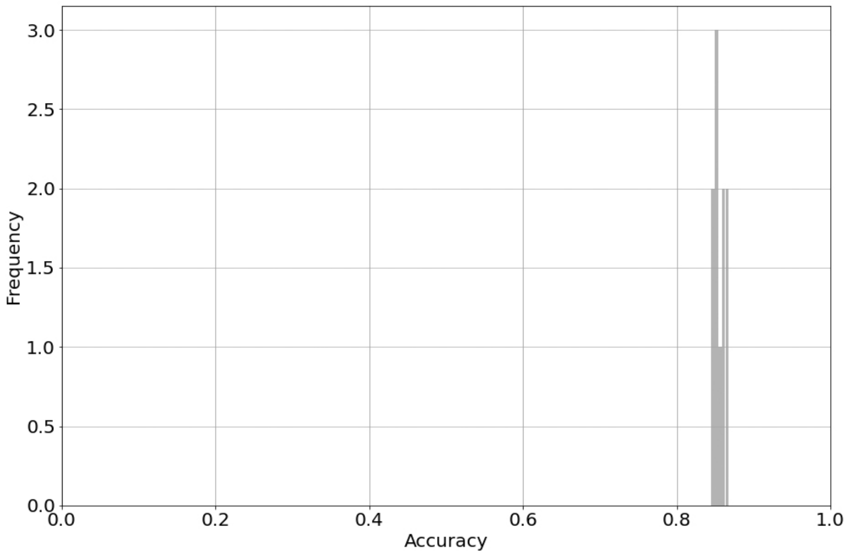

Fig. 1. The distribution of accuracy of default prediction for training data formed under 'Year-before-approach' (model with financial data only)

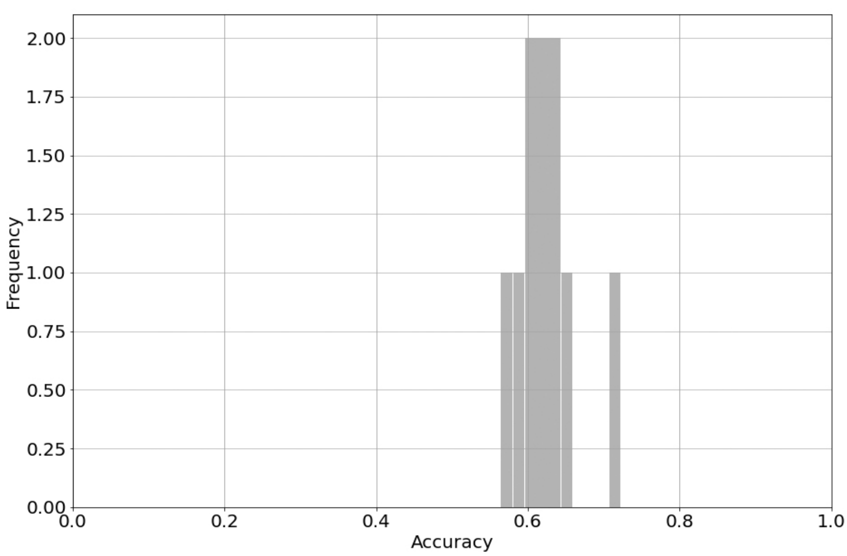

Fig. 2. The distribution of accuracy of default prediction for test data formed under 'Year-before-approach' (model with financial data only)

Figure 3 shows the confusion matrices for 10 test samples formed under 'Year-before-approach'. The prediction was made using financial data only. It is seen that the sensitivity (the ability of the algorithm to identify defaulted firms) is more than 65% for every case, but the specificity (the ability of the algorithm to identify healthy peers) is low and even less than 50% for some of the samples (which means that the algorithm makes mistakes more often than right predictions). These results suggest that the model is inapplicable for default prediction.

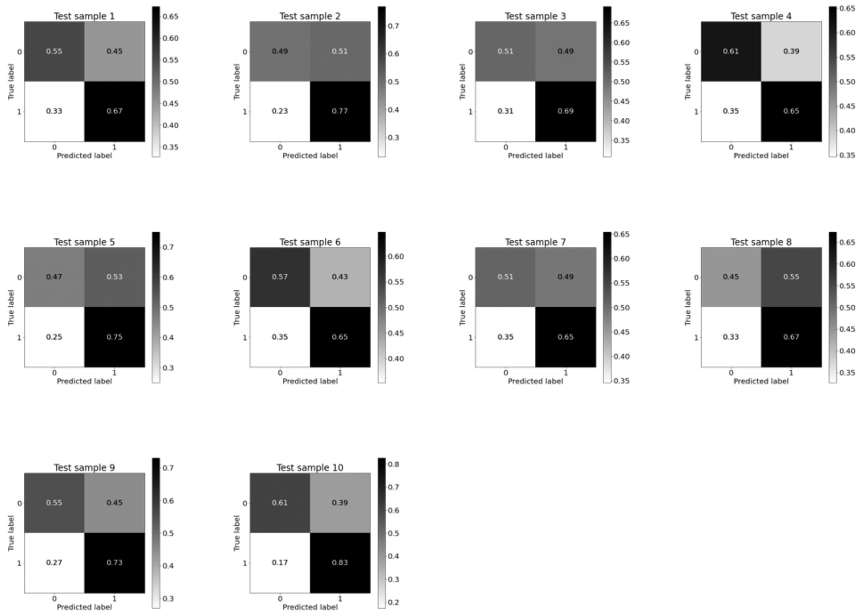

Fig. 3. Confusion matrices for test samples under 'Year-before-approach' (model with financial data only), 1 — defaulted, 0 — healthy peers

4.2. 'Year-before-approach', model with financial and non-financial data

A higher accuracy was achieved while training Random Forest algorithm on 10 training samples, formed under 'Year-before-approach' using financial and non-financial data together. The mean accuracy of default prediction on the training samples is 86.0% (see Figure 4), the mean accuracy on the test samples is 70.6% (see Figure 5). The average area under ROC curve is 0.95 for training data and 0.78 for test data.

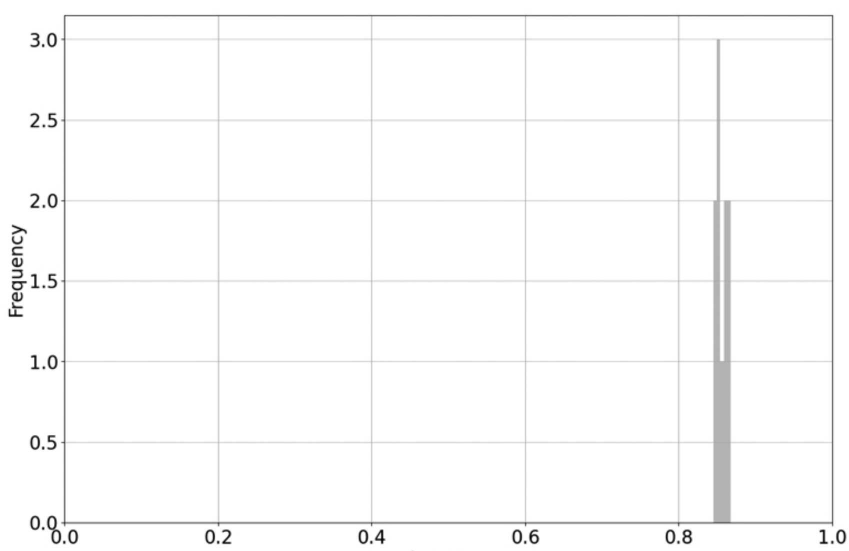

Fig. 4. The distribution of accuracy of default prediction for training data formed under 'Year-before-approach' (model with financial and non-financial data)

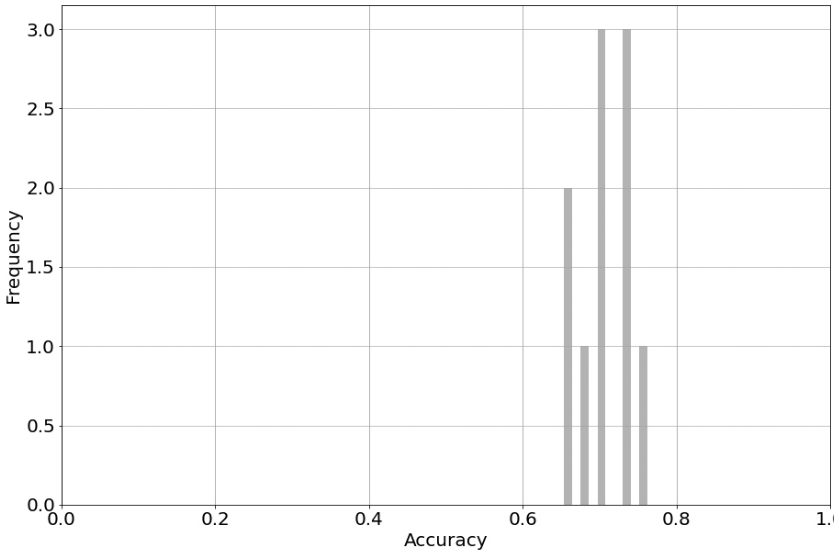

Fig. 5. The distribution of accuracy of default prediction for test data formed under 'Year-before-approach' (model with financial and non-financial data)

Figure 6 shows the confusion matrices for 10 test samples formed under 'Year-before-approach' using financial and non-financial data together. In half of the cases, it is seen that both sensitivity and specificity are around 70%. Despite such values are not very high, these results suggest that the model, trained on both financial and non-financial data is a way more accurate than a model, trained on financial data only.

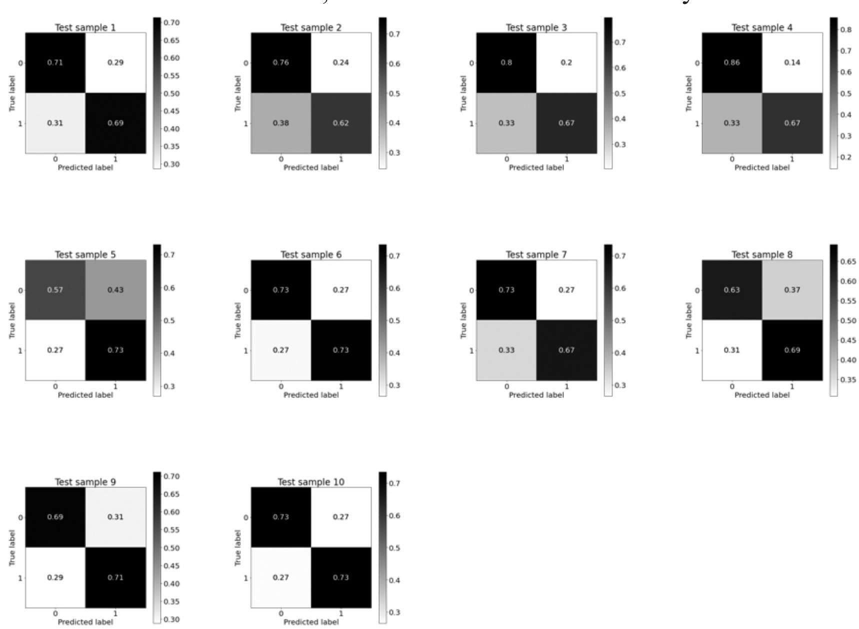

Fig. 6. Confusion matrices for test samples under 'Year-before-approach' (model with financial and non-financial data), 1 — defaulted, 0 — healthy peers

The value of non-financial variables in classifying the default firms is proven by high importance metric (see Figure 7). The two most 'important' features for classification are the sum and the number of legal claims against the firm during 1 year before the theoretical forecast date.

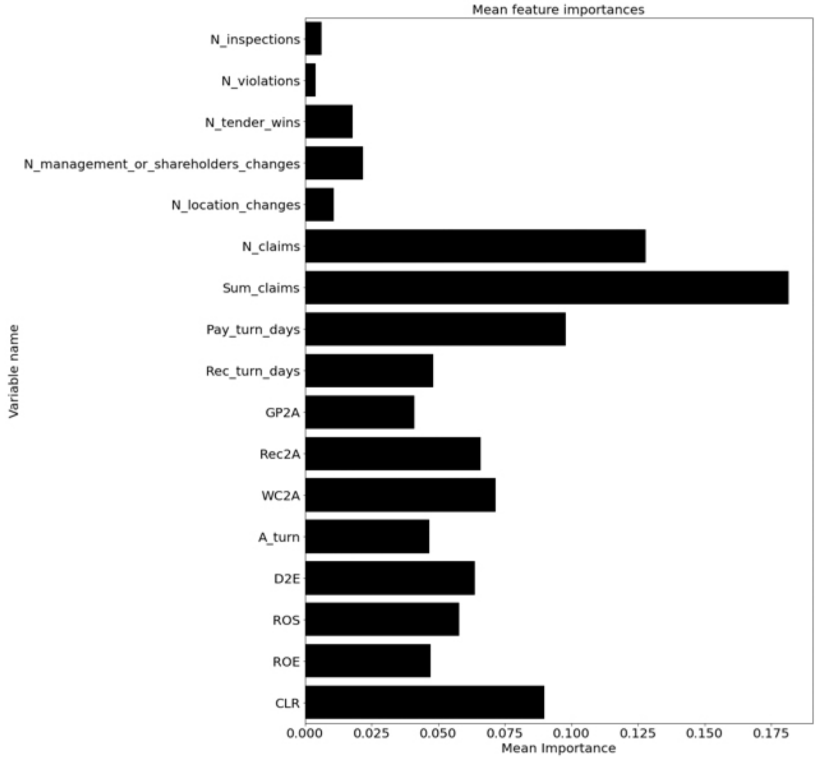

Fig. 7. Feature importances under 'Year-before-approach' (model with financial and non-financial data)

4.3. 'Last-available-financial-report-approach', model with financial data only

Then I trained Random Forest algorithm on 10 training samples formed under 'Last-available-financial-report-approach' using only financial data. The mean accuracy of default prediction on the training samples in this case turned to be 87.5% (see Figure 8), the mean accuracy on the test samples — 64.5% (see Figure 9). The average area under ROC curve is 0.96 for training data and 0.71 for test data.

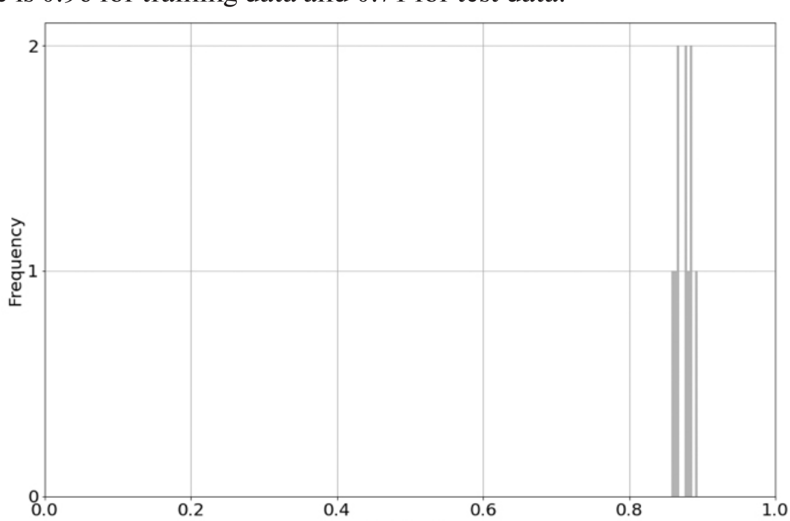

Fig. 8. The distribution of accuracy of default prediction for training data formed under 'Last-available-financial-report-approach' (model with financial data only)

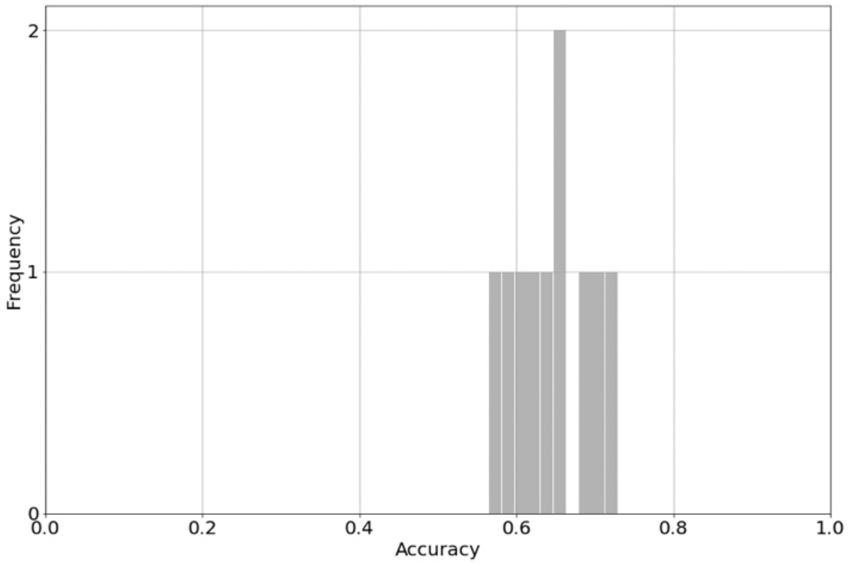

Fig. 9. The distribution of accuracy of default prediction for test data formed under 'Last-available-financial-report-approach' (model with financial data only)

Figure 10 shows the confusion matrices for 10 test samples formed under 'Last-available-financial-report-approach' using only financial data. Seems that the 'Last-available-financial-report-approach' gives more accurate classification results than the 'Year-before-approach' using only financial data, but still in most of the cases the classification is not sufficiently accurate.

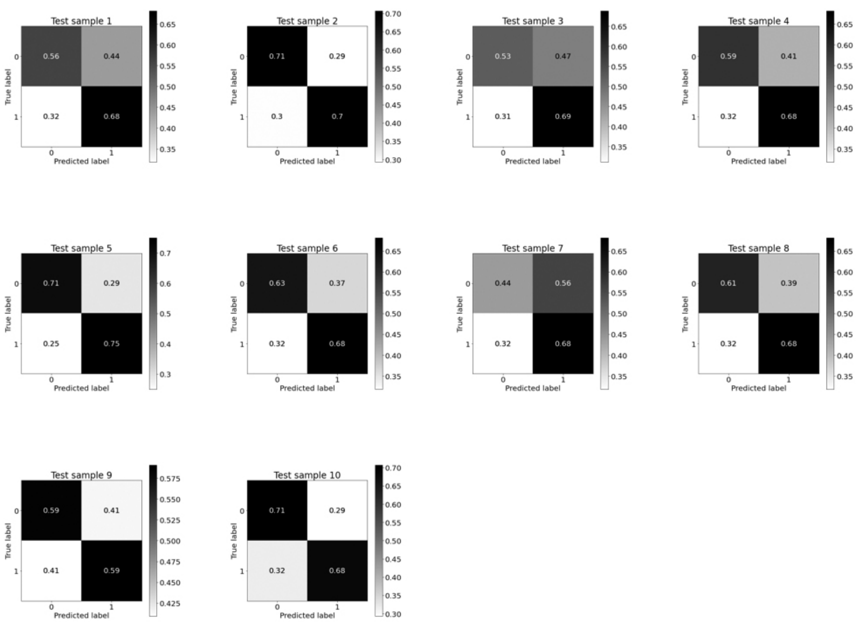

Fig. 10. Confusion matrices for test samples under 'Last-available-financial-report-approach' (model with financial data only), 1 — defaulted, 0 — healthy peers

4.4. 'Last-available-financial-report-approach', model with financial and non-financial data

As in the case of 'Year-before-approach', the model, based on both financial and non-financial data provided higher accuracy. The mean accuracy of default prediction on the training samples is 88.6% (see Figure 11), the mean accuracy on the test samples is 72.3% (see Figure 12). The average area under ROC curve is 0.96 for training data and 0.82 for test data.

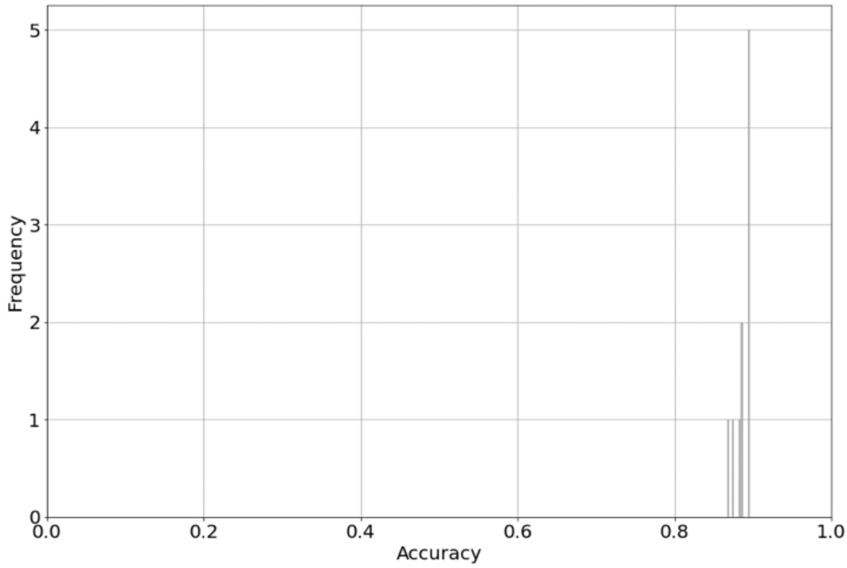

Fig. 11. The distribution of accuracy of default prediction for training data formed under 'Last-available-financial-report-approach' (model with financial and non-financial data)

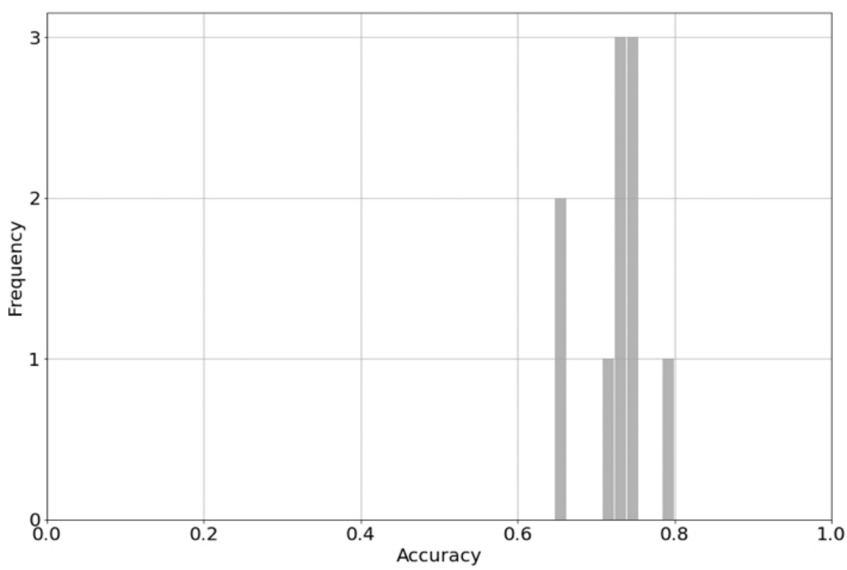

Fig. 12. The distribution of accuracy of default prediction for test data formed under 'Last-available-financial-report-approach' (model with financial and non-financial data)

Figure 13 shows the confusion matrices for 10 test samples formed under 'Last-available-financial-report-approach' using financial and non-financial data together. In the majority of cases, both sensitivity and specificity are higher than 70%, so the model seems to be the best one among all models built within this research.

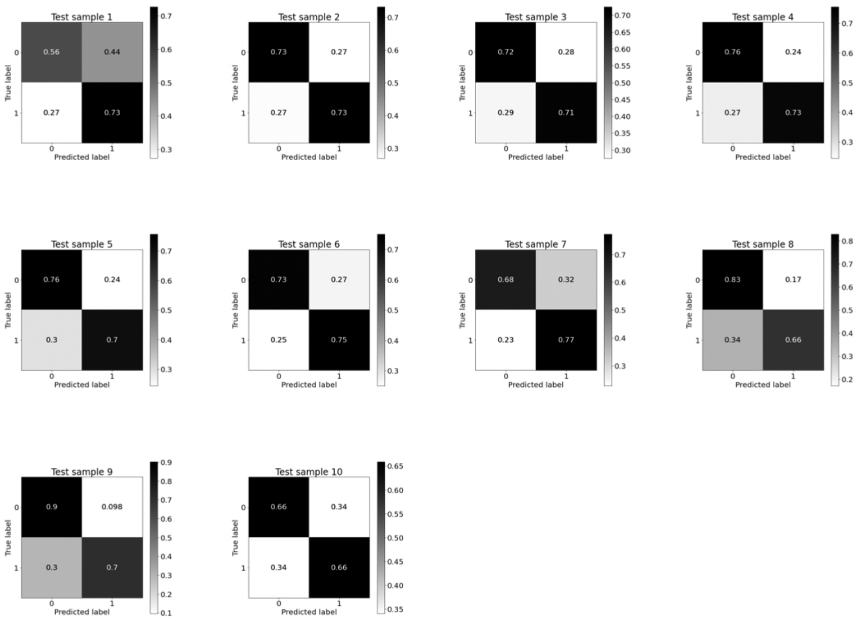

Fig. 13. Confusion matrices for test samples under 'Last-available-financial-report-approach' (model with financial and non-financial data), 1 — defaulted, 0 — healthy peers

The feature importance data again suggest that the sum and the number of legal claims are the most 'important' features for the classification, however other non-financial data seem to be less 'important' than financial variables (see Figure 14).

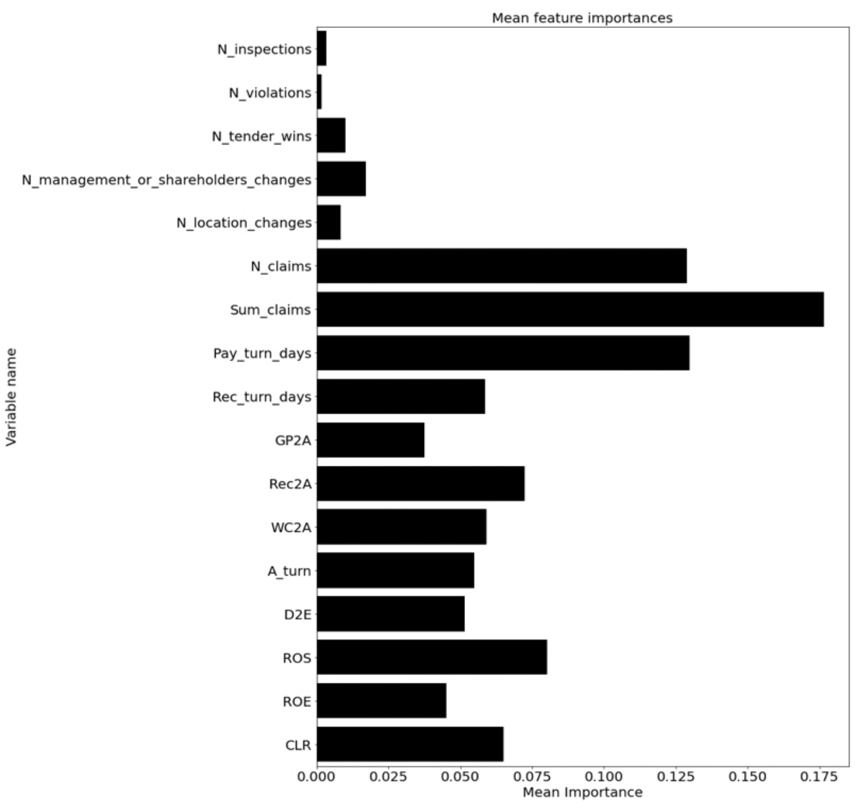

Fig. 14. Feature importances under 'Last-available-financial-report-approach' (model with financial and non-financial data)

4.5. Models comparison

2 models, based on non-financial data only were fitted as well, the classification accuracy turned to be low, however, higher than in the case of the models based on financial data only. Thus, based on the provided results one can conclude that non-financial data, namely legal claims related information, can improve the quality of default prediction if used along with financial ratios. Table 7 shows the mean prediction quality metrics on the test samples for all fitted models.

Table 7

Mean prediction quality metrics on the test data

	'Year-before-approach'			*'Last-available-financial-report-approach'*		
Metric	*Financial data only*	*Financial and non-financial*	*Non-financial data only*	*Financial data only*	*Financial and non-financial*	*Non-financial data only*
Accuracy	62%	71%	66%	65%	**72%**	71%
Sensitivity	71%	69%	52%	68%	**71%**	63%
Specificity	53%	72%	81%	61%	**73%**	80%
AUC ROC	0.68	0.78	0.69	0.71	**0.82**	0.75

5. Conclusions

This study aimed to test whether the use of non-financial data along with financial data can significantly improve the accuracy of default prediction. I chose 10 financial ratios, which represent profitability, liquidity, solvency and turnover ratios, and used them to create the baseline models based on financial data only. Also, 7 non-financial indicators were chosen, representing the information related to legal claims, inspections, tenders and location, CEO and shareholders changes.

The models were fitted on a training samples, drawn from more than 2000 auto repair firms. A special attention was paid to the proper choice of the theoretical forecast date (the date, for which the values of the predictors are calculated), and I used two approaches to form the training and testing samples. The first approach was to take the values of the predictors, available exactly one year before the default ('Year-before-approach'). In this case the accuracy of prediction reflects the estimate of the ability to predict default exactly one year before it. The second approach was to take the values, available on the most recent reporting date ('Last-available-financial-report-approach'). In this case the accuracy of prediction reflects the estimate of the ability to predict default in the period starting from the date of financial reporting.

The results show that, firstly, financial ratios do not seem to be enough to predict default for auto repair private firms in an emerging economy. The accuracy is low (not more than 65%). Secondly, using non-financial variables along with financial data can significantly improve the accuracy of prediction (almost 10 percentage points increase). Thirdly, the most important non-financial variables turned to be the value and the number of legal claims against the firm during one year before the theoretical forecast date. Finally, the accuracy of prediction is higher under 'Last-available-financial-report-approach', probably because of the theoretical forecast date is closer to the real default date.

The conclusions of this study are significant, firstly, for the credit institutions, which work with private firms in emerging economies. When there is a concern about the relevance of submitted financial reporting, it can be beneficial to take into account the available non-financial data. Secondly, the findings are important for the business. The check of the reliability of the counterparty includes the credit risk estimation (which is, actually, the probability of default estimation). It is not always possible to check the financial ratios of a counterparty, but the open access non-financial data is always available for any private firm. Finally, the findings of this study can be a signal for the bias in the financial ratios of auto repair firms, which is an important conclusion for the policy makers.

References

Afanasev V. Default Prediction Model for Emerging Capital Market Service Companies // Journal of Corporate Finance Research. 2023. Vol. 17. Iss. 1. P. 64–77.

Ahmadpour Kasgari A., et al. Prediction of Bankruptcy Iranian Corporations Through Artificial Neural Network and Probit-based Analyses // Neural Comput & Applic. 2013. Vol. 23. Iss. 3. P. 927–936.

Altman E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy // The Journal of Finance. 1968. Vol. 23. Iss. 4. P. 589–609.

Altman E., Sabato G., Wilson N. The Value of Non-financial Information in SME Risk Management // The Journal of Credit Risk. 2010. Vol. 6. P. 95–127.

Andrade G., Kaplan S. N. How Costly is Financial (Not Economic) Distress? Evidence from Highly Leveraged Transactions that Became Distressed // The Journal of Finance. 1998. Vol. 53. Iss. 5. P. 1443–1493.

Bagherpour Velashani M. A., Jadaan Sallal F. F., Salehi M. COVID-19 Pandemic and Fraudulent Financial Reporting Motives in Emerging Markets: A Comparative Study // IJBDS. 2022. Vol. 14. Iss. 1.

Barboza F., Kimura H., Altman E. Machine Learning Models and Bankruptcy Prediction // Expert Systems with Applications. 2017. Vol. 83. P. 405–417.

Bazzana F., Bee M., Hussin Adam Khatir A. A. Machine Learning Techniques for Default Prediction: an Application to Small Italian Companies // Risk Manag. 2024. Vol. 26. Iss. 1. P. 1.

Beaver W. H. Financial Ratios As Predictors of Failure // Journal of Accounting Research. 1966. Vol. 4. P. 71–111.

Bhimani A., Gulamhussen M. A., Lopes S. da R. The Role of Financial, Macroeconomic, and Non-financial Information in Bank Loan Default Timing Prediction // European Accounting Review. 2013. Vol. 22. Iss. 4. P. 739–763.

Blanco-Oliver A., Irimia-Dieguez A., Oliver-Alfonso M., Wilson N. Improving Bankruptcy Prediction in Micro-entities by Using Nonlinear Effects and Non-financial Variables // Finance a Uver — Czech Journal of Economics and Finance. 2015. Vol. 65. Iss. 2. P. 144–166.

Boubaker S., Cellier A., Manita R., Saeed A. Does Corporate Social Responsibility Reduce Financial Distress Risk? // Economic Modelling. 2020. Vol. 91(8). P. 835–851.

Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. Iss. 1. P. 5–32.

Brown I., Mues C. An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets // Expert Systems with Applications. 2012. Vol. 39. Iss. 3. P. 3446–3453.

Cao Y. et al. A Two-stage Bayesian Network Model for Corporate Bankruptcy Prediction // International Journal of Finance & Economics. 2020. Vol. n/a. Iss. n/a.

Gruszczyński M. Financial Distress of Companies in Poland // Department of Applied Econometrics, Warsaw School of Economics. Working Papers. 2004. Vol. 10.

Hogan C., Rezaee Z., Riley J., Velury U. Financial Statement Fraud: Insights from the Academic Literature // Auditing: A Journal of Practice & Theory. 2008. Vol. 27. Iss. 2. P. 231–252.

Jaki A., Ćwięk W. Bankruptcy Prediction Models Based on Value Measures // Journal of Risk and Financial Management. 2021. Vol. 14. Iss. 1. P. 6.

Kumar P., Ravi V. Bankruptcy Prediction in Banks and Firms Via Statistical and Intelligent Techniques — A review // European Journal of Operational Research. 2007. Vol. 180. Iss. 1. P. 1–28.

Laitinen E. K. Assessing Viability of Finnish Reorganization and Bankruptcy Firms // European Journal of Law and Economics. 2011. Vol. 31. Iss. 2. P. 167–198.

Lin F., Liang D., Chu W.-S. The role of Non-financial Features Related to Corporate Governance in Business Crisis Prediction // Journal of Marine Science and Technology. 2010. Vol. 18. Iss. 4.

Lugovskaya L. Predicting Default of Russian SMEs on the Basis of Financial and Non-financial Variables // Journal of Financial Services Marketing. 2010. Vol. 14. Iss. 4. P. 301–313.

Lukason O., Hoffman R. Firm Bankruptcy Probability and Causes: An Integrated Study // International Journal of Business and Management. 2014. Vol. 9. P. 80–91.

Mai F., Tian S., Lee C., Ma L. Deep learning models for bankruptcy prediction using textual disclosures // European Journal of Operational Research. 2019. Vol. 274. Iss. 2. P. 743–758.

Mselmi N., Lahiani A., Hamza T. Financial Distress Prediction: The Case of French Small and Medium-sized Firms // International Review of Financial Analysis. 2017. Vol. 50. P. 67–80.

Muñoz-Izquierdo N., Laitinen E., Camacho-Miñano M., 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. Vol. 31. Iss. 1. P. 65–97.

Ohlson J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy // Journal of Accounting Research. 1980. Vol. 18. Iss. 1. P. 109–131.

Schwartz K. B., Menon K. Auditor Switches by Failing Firms // The Accounting Review. 1985. Vol. 60. Iss. 2. P. 248–261.

Shumway T. Forecasting Bankruptcy More Accurately: A Simple Hazard Model // Journal of Business. 2001. Vol. 74. Iss. 1. P. 101.

Uddin M. S. et al. Leveraging Random Forest in Micro-Enterprises Credit Risk Modelling for Accuracy and Interpretability // International Journal of Finance & Economics. 2022. Vol. 27. Iss. 3. P. 3713–3729.

Wilson N., Ochotnický P., Káčer M. Creation and Destruction in Transition Economies: The SME Sector in Slovakia // International Small Business Journal. 2016. Vol. 34. Iss. 5. P. 579–600.

Xie C., Luo C., Yu X. Financial Distress Prediction Based on SVM and MDA Methods: The case of Chinese Listed Companies // Quality and Quantity. 2011. Vol. 45. P. 671–686.

Zhao Y., Lin D. Prediction of Micro- and Small-Sized Enterprise Default Risk Based on a Logistic Model: Evidence from a Bank of China // Sustainability. 2023. Vol. 15. Iss. 5. P. 4097.

Zmijewski M. E. Methodological Issues Related to the Estimation of Financial Distress Prediction Models // Journal of Accounting Research. 1984. Vol. 22. P. 59.

References

Afanasev V. Default Prediction Model for Emerging Capital Market Service Companies. *Journal of Corporate Finance Research*, 2023, vol. 17, N 1, pp. 64–77.

Ahmadpour Kasgari A. et al. Prediction of bankruptcy Iranian corporations through artificial neural network and Probit-based analyses. *Neural Comput & Applic*, 2013, vol. 23, N 3, pp. 927–936.

Altman E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 1968, vol. 23, N 4, pp. 589–609.

Altman E., Sabato G., Wilson N. The value of non-financial information in SME risk management. *The Journal of Credit Risk*, 2010, vol. 6, pp. 95–127.

Andrade G., Kaplan S. N. How Costly is Financial (Not Economic) Distress? Evidence from Highly Leveraged Transactions that Became Distressed. *The Journal of Finance*, 1998, vol. 53, N 5, pp. 1443–1493.

Bagherpour Velashani M. A., Jadaan Sallal F. F., Salehi M. COVID-19 Pandemic and Fraudulent Financial Reporting Motives in Emerging Markets: A Comparative Study. *IJBDS*, 2022, vol. 14, N 1.

Barboza F., Kimura H., Altman E. Machine learning models and bankruptcy prediction. *Expert Systems with Applications*, 2017, vol. 83, pp. 405–417.

Bazzana F., Bee M., Hussin Adam Khatir A. A. Machine learning techniques for default prediction: an application to small Italian companies. *Risk Manag*, 2024, vol. 26, N 1, pp. 1.

Beaver W. H. Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 1966, vol. 4, pp. 71–111.

Bhimani A., Gulamhussen M. A., Lopes S. da R. The Role of Financial, Macroeconomic, and Non-financial Information in Bank Loan Default Timing Prediction. *European Accounting Review*, 2013, vol. 22, N 4, pp. 739–763.

Blanco-Oliver A., Irimia-Dieguez A., Oliver-Alfonso M., Wilson N. Improving bankruptcy prediction in micro-entities by using nonlinear effects and non-financial variables. *Finance a Uver — Czech Journal of Economics and Finance*, 2015, vol. 65, N 2, pp. 144–166.

Boubaker S., Cellier A., Manita R., Saeed A. Does corporate social responsibility reduce financial distress risk? *Economic Modelling*, 2020, vol. 91(8), pp. 835–851.

Breiman L. Random Forests. *Machine Learning*, 2001, vol. 45, N 1, pp. 5–32.

Brown I., Mues C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, 2012, vol. 39, N 3, pp. 3446–3453.

Cao Y. et al. A two-stage Bayesian network model for corporate bankruptcy prediction. *International Journal of Finance & Economics*, 2020, vol. n/a, N n/a.

Gruszczyński M. Financial Distress of Companies in Poland. *Department of Applied Econometrics, Warsaw School of Economics, Working Papers*, 2004, vol. 10.

Hogan C., Rezaee Z., Riley J., Velury *U.* Financial Statement Fraud: Insights from the Academic Literature. *Auditing: A Journal of Practice & Theory*, 2008, vol. 27, N 2, pp. 231–252.

Jaki A., Ćwięk W. Bankruptcy Prediction Models Based on Value Measures. *Journal of Risk and Financial Management*, 2021, vol. 14, N 1, pp. 6.

Kumar P., Ravi V. Bankruptcy prediction in banks and firms via statistical and intelligent techniques — A review. *European Journal of Operational Research*, 2007, vol. 180, N 1, pp. 1–28.

Laitinen E. K. Assessing viability of Finnish reorganization and bankruptcy firms. *Eur J Law Econ*, 2011, vol. 31, N 2, pp. 167–198.

Lin F., Liang D., Chu W.-S. The role of non-financial features related to corporate governance in business crisis prediction. *Journal of Marine Science and Technology*, 2010, vol. 18, N 4.

Lugovskaya L. Predicting default of Russian SMEs on the basis of financial and non-financial variables. *Journal of Financial Services Marketing*, 2010, vol. 14, N 4, pp. 301–313.

Lukason O., Hoffman R. Firm Bankruptcy Probability and Causes: An Integrated Study. *International Journal of Business and Management*, 2014, vol. 9, pp. 80–91.

Mai F., Tian S., Lee C., Ma L. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, 2019, vol. 274, N 2, pp. 743–758.

Mselmi N., Lahiani A., Hamza T. Financial distress prediction: The case of French small and medium-sized firms. *International Review of Financial Analysis*, 2017, vol. 50, pp. 67–80.

Muñoz-Izquierdo N.,. Laitinen E., Camacho-Miñano M., 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, vol. 31, N 1, pp. 65–97.

Ohlson J. A. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 1980, vol. 18, N 1, pp. 109–131.

Schwartz K. B., Menon K. Auditor Switches by Failing Firms. *The Accounting Review*, 1985, vol. 60, N 2, pp. 248–261.

Shumway T. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *Journal of Business*, 2001, vol. 74, N 1, pp. 101.

Uddin M. S. et al. Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability. *International Journal of Finance & Economics*, 2022, vol. 27, N 3, pp. 3713–3729.

Wilson N., Ochotnický P., Káčer M. Creation and destruction in transition economies: The SME sector in Slovakia. *International Small Business Journal*, 2016, vol. 34, N 5, pp. 579–600.

Xie C., Luo C., Yu X. Financial distress prediction based on SVM and MDA methods: The case of Chinese listed companies. *Quality and Quantity*, 2011, vol. 45, pp. 671–686.

Zhao Y., Lin D. Prediction of Micro- and Small-Sized Enterprise Default Risk Based on a Logistic Model: Evidence from a Bank of China. *Sustainability*, 2023, vol. 15, N 5, pp. 4097.

Zmijewski M. E. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research*, 1984, vol. 22, pp. 59.

Статья поступила в редакцию 20 февраля 2024 г.
Статья рекомендована в печать 15 мая 2024 г.