Bank of Russia

Probability of default model using transaction data of Russian companies

WORKING PAPER SERIES

No. 97 / June 2022

Andrey Shevelev
Gleb Buzanov

Andrey Shevelev
Bank of Russia. Email: ShevelevAA@cbr.ru

Gleb Buzanov
Bank of Russia. Email: BuzanovGS@cbr.ru

The authors would like to thank Sergei Seleznev, Sergei Glechikov, and the participants of the internal research seminar at the Bank of Russia for their helpful comments and suggestions. All errors and omissions are ours.

The Bank of Russia Working Paper Series is anonymously refereed by members of the Bank of Russia Research Advisory Board and external reviewers.

Cover image: Shutterstock.com

Address: 12 Neglinnaya Street, Moscow, 107016
Tel.: +7 495 771-91-00, +7 495 621-64-65 (fax)
Website: www.cbr.ru

Abstract

The purpose of this study is to test the usefulness of transaction data from the Bank of Russia Payment System (BRPS), to predict the default probabilities of Russian companies. To fulfil this purpose, we build probability of default models for each industry group using machine learning methods based on annual accounting data. Thereafter, we add features generated from transaction data to the models and improve their forecast quality according to the ROC AUC metric.

Additionally, we train our probability of default models for each industrial group using a Random Forest based only on BRPS data. The forecast quality of this is a little worse on average according to the ROC AUC metric, but these estimates can be obtained at least three months earlier than estimates using annual accounting statements.

Our results confirm that BRPS transaction data are useful for improving the quality of forecasting the default probabilities of Russian companies. In addition, the Random Forest feature importance shows that the main sources of this additional information are payroll taxes and social payments.

Key words: probability of default model, transactional data, Random Forest, logistic regression

JEL codes: C53, C5, E44

Contents

1. Introduction

Ensuring uninterrupted operation of the financial sector and increasing its stability is one of the key functions of the Bank of Russia. For this purpose, the financial system is constantly monitored and macroprudential tools are applied. To this end, the Bank of Russia considers, among other things, indicators of the default probabilities of Russian companies contained in portfolios of commercial banks.

The relevance of a prompt and accurate assessment of the state of companies in a bank's loan portfolio and their probability of default is increasing, especially in the conditions of the modern economic system. If forecasts for a borrower's financial condition are overly conservative, the bank may classify healthy cases as bad, having to build excessive loan reserves. As a result, the bank would have suboptimal capital structures and a reduced credit supply. On the contrary, excessive optimism in the model will lead to a deficit of reserves, which may entail the risk of bank default and, consequently, problems in the financial system.

Most general models for predicting company defaults are based on accounting data and information on loans, and some works take macroeconomic and financial indicators into account. The disadvantages of these approaches are the infrequent publication of data (annual accounting statements), delays in the publication of data (three months or more), and the lack of accounting for interaction between economic agents. Data from the Bank of Russia Payment System make it possible to solve these problems.

The Bank of Russia Payment System (BRPS) is 'a systemically important payment system that plays a key role in the implementation of monetary and budgetary policy. It also plays a central part in settling payments by financial market participants, including most interbank payments.'[1] In 2015, 1,398.5 million payment transactions with a value of 1,356.5 trillion rubles were made through the BRPS. The average daily volume of payments processed through the BRPS in 2015 was 5.5 trillion rubles and the average daily number of payments was 5.6 million.[2]

The creation of a state-of-the-art model for predicting the probability of defaults is a labour-intensive task and processing and applying such a large amount of transactional data would require even more time and computing resources. Therefore, to begin with, we do not set ourselves the task of building the best model. The purpose of this study is to test the usefulness of transaction data from the BRPS to predict the default probabilities of Russian companies.

To fulfil the purpose of this study, in the initial stage, we build Random Forest models to predict company default probabilities in each industry group using standard annual accounting data. We then add features generated from transaction data to the models. To verify the results, we also train Logistic Regression models with L2 regularisation and weighted likelihood functions. In addition, we train Random Forest models using only

[1] The Bank of Russia Payment System https://www.cbr.ru/eng/Psystem/payment_system/

[2] World Bank; International Finance Corporation. 2016. Russian Federation Financial Sector Assessment Program: Financial Infrastructure Technical Note. World Bank, Washington, DC. © World Bank. https://openknowledge.worldbank.org/handle/10986/25066 License: CC BY 3.0 IGO.

transaction data to show that these data allow us to obtain predictions of the probability of defaults at least three months earlier than those of models based on accounting data.

The paper is structured as follows. The next section includes a brief overview of the literature on the prediction of default probabilities. Section 3 contains a description of the data that were used in the models. Section 4 gives details on the machine learning models used, the hyperparameter optimisation techniques used in finding the optimal model architecture, and techniques for dealing with unbalanced samples. Section 5 describes the model results estimated for each industry group and a robustness check. The conclusion section summarises the main results of the work.

2. Literature review

The work of Beaver (1966) laid the foundation for the modern literature on the forecasting of insolvency, which Beaver (1968) later explored in more detail. The most popular model in this class is that presented by Altman (1968). These methods were based on a multivariate framework with extensive use of multivariate discriminant analysis (MDA) models and the construction of a ‘Z-score’ to divide borrowers into potential bankrupts and non-bankrupts. However, criticism of violations of the statistical assumptions underlying the MDA approach led researchers in the 1980s to focus their efforts on the development of conditional probability models. The most popular of these is the logit model of, e.g., Ohlson (1980). The work of Odom and Sharida (1990) was one of the first to use a neural network consisting of several hidden layers in predicting bankruptcies. Financial coefficients used in Altman's model were taken as input data.

The study ‘The performance of insolvency prediction and credit risk models in the UK: A comparative study’ by Jackson and Wood (2013) shows that 25 different methods developed over the past five decades provide different results and have different predictive abilities. In terms of forecast accuracy, each of these models outperforms earlier models.

In the last decade, Russian researchers have begun to pay attention to the construction of probability of default models for Russian companies. Demeshev and Tikhonova (2014) compare approaches to modelling the critical financial situation of medium and small non-public Russian companies in four industries (manufacturing, real estate, wholesale and retail trade, and construction) using financial and non-financial indicators for the years 2011–2012. The study is based on the analysis of annual data in the financial statements (balance sheet and income statement) of non-public Russian companies in the RUSLANA database for the period 2011–2012. The paper provides a list of the financial and non-financial indicators used. Of the algorithms selected by the authors for comparison (linear discriminant analysis (LDA), quadratic DA, DA of mixtures of distributions, the method of classification trees, and Random Forest), the Random Forest algorithm showed the greatest predictive power. The authors also note that among the non-financial indicators, the industry, the federal district, and the age of the enterprise have a strong influence. The size of the enterprise is less important, and its organisational form turned out to be practically insignificant. Among the financial indicators, the most important were the ratios of profitability, financial leverage, and liquidity.

A paper by Mogilat (2015) investigates company bankruptcy in Russia and identifies and analyses the main trends and structural characteristics of bankrupt and non-bankrupt companies in the 2007–2014 period. It shows that the main factors for predicting the probability of default are the net return on assets, the turnover of total assets, the ratio of the company's net accounts payable to its total assets, and return on assets in the industry. These factors remain consistently significant even with changes in the set of control factors and the variation of the sample. A later article by Mogilat (2019) proposes an econometric approach based on logistic regression to the assessment of the risks to the financial stability of Russian industrial companies that takes both global experience with such studies and the particularities of Russian data into account. Data from the financial statements of Russian companies from 2006 to 2016 are used. After the filtering procedure, the database contains an average of about 97,000 companies per year. To identify bankruptcies, data from 2007 to 2017 are analysed. The factors for modelling financial sustainability are selected based on global practice and the approach previously proposed by Mogilat (2015). The paper also discusses the problem of imbalanced data and methods for solving it. The author uses a weighted likelihood function to solve this problem.

Another paper, by Karminsky and Burekhin (2019), compares algorithms such as logit and probit models, classification trees, random forests, and neural networks to predict the probability of default of Russian companies in the construction industry. Particular attention is paid to the peculiarities of building machine learning models, the impact of data imbalance on the predictive power of models, and the analysis of ways to combat it, as well as to the analysis of the influence of non-financial factors. This paper uses indicators based on the public financial statements of companies from 2011 to 2017. Based on these financial statements, financial indicators that reflect the economic activities of the company are calculated. These can be briefly described as indicators of profitability, liquidity, business activity, and financial stability.

Since the 1970s, financial indicators based on accounting statements have been an important source of data for the construction of probability of default models. However, the information used to build such models does not take into account the volatility of a company's performance during the period analysed, which is a reason to criticise this approach. The Merton model, whose theoretical foundations are the basis of the KMV model, is a model that explains the default of a company by a fall in the value of its assets. Merton (1974) proposed a model in which the common stock issued by a company is interpreted as an option on that company's assets. The main advantage of the options approach is that it allows the derivation of the probability of default from the price values observed in the market. At the same time, this model has a significant disadvantage of particular relevance to Russia: the limited database of private enterprises that have shares in circulation on the market.

To take into account the fluctuation of the financial indices of a company during a year and to consider companies whose shares are not traded on the market, we can use payment transaction data. As of the time of writing, we have found no studies of default forecasting on transaction data for Russian companies, although there are a number of works that address related issues. An article by Babaev et al. (2019) describes the use of

deep learning models based on transaction data to assess the creditworthiness of retail banking customers. The authors use Embedding-Transactional RNN (ET-RNN, the network architecture is presented in the article) on customer transaction data (payment amount, card type, date and time of payment, country, currency, and transfer type). The training sample is composed of the transactions of approximately 740,000 clients. The total number of transactions is about 200 million (with around 800 transactions per client). Consumer loan defaults during the year are used as the target variable. The authors use logistic regression (with 400 factors) and LightGBM (with 7,000 generated factors) as baseline models. As a result, the authors show that the Logistic Regression model has an ROC AUC metric of 0.78, the LightGBM model has 0.81, and the ET-RNN model increases the metric to 0.83 points. It should be noted that in contrast to the classical methods, which largely depend on generated factors, the ET-RNN method did not require the manual generation of the factors.

Returning to the purpose of our study, in order to test the usefulness of BRPS data, we focus on the Random Forest model as a baseline and use the Logistic Regression model as a robustness check. As variables for building the probability of default models, we take a standard set of financial indicators described in more detail in the next section.

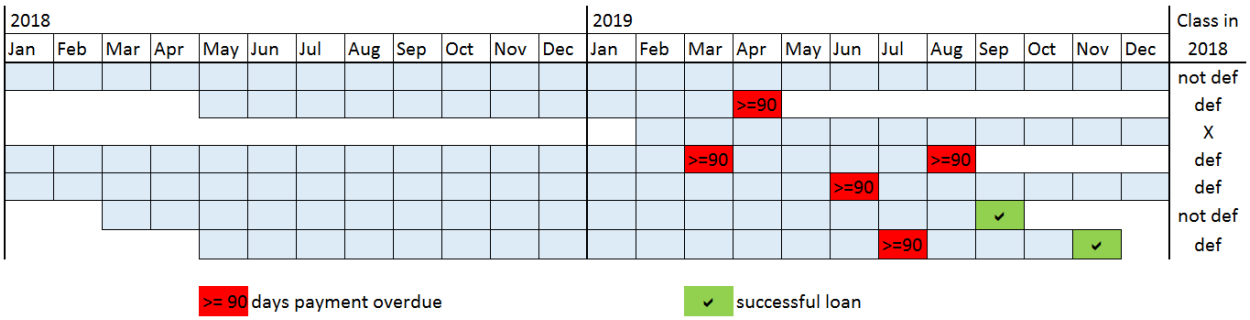
3. Data

Before predicting probabilities of default, it is necessary to define company default. We define a default event as a case of a payment overdue by more than 90 days according to the Basel II (III) Internal ratings-based (IRB) methodology (BCBS 2017). The default date is determined using information on overdue debt based on data from credit bureaus before 2018, and from Reporting Form 0409303 after. This form is submitted monthly, so we examine each period and look at the overdue debt. If the duration is more than 90 days, we mark that the company has a default. For a more detailed explanation of the definition of default, see Figure 1 (the ‘X’ in the last column means that company is not included in our sample).

Accounting data from 2012 to 2018 were obtained from the website of the Federal State Statistics Service.[3] Since our goal is to check the usefulness of the transaction data from the BRPS that are available from 2015, we use accounting data to calculate models from 2015. These accounting data were converted into financial indicators for modelling (see Table 1 in the Appendix for more details).

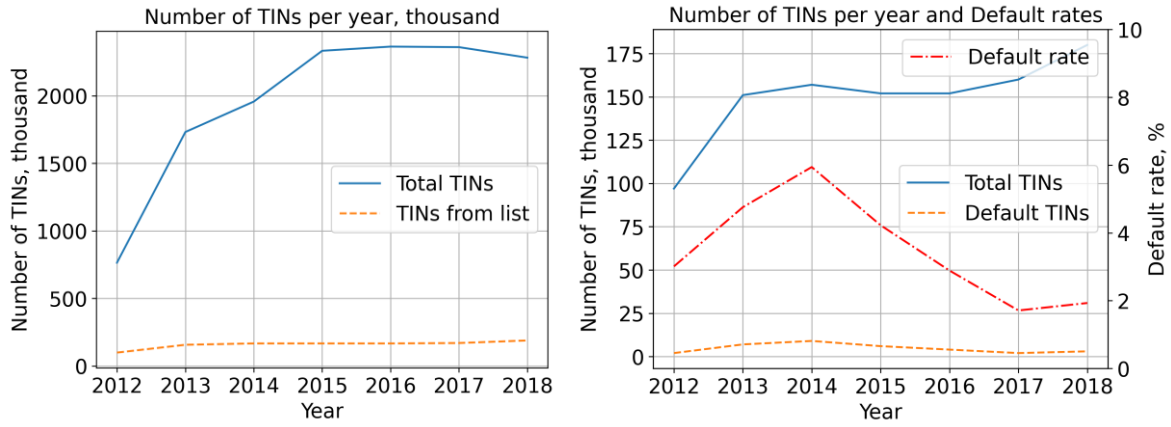
[3] Rosstat website: https://rosstat.gov.ru/

Figure 1. Definition of company default



The initial data sample ranged from about 700,000 unique tax identification numbers (TIN) in 2012 to more than 2,000,000 unique identification numbers in 2018. Companies that had a loan in the year following the reporting period were included in the dataset (called 'TINs from list' in Figure 2). The default rates across all companies included ranged from 2% to 6% per year, which shows the imbalance of classes in the data.

Figure 2. Number of TINs per year and TINs under our consideration with default rates

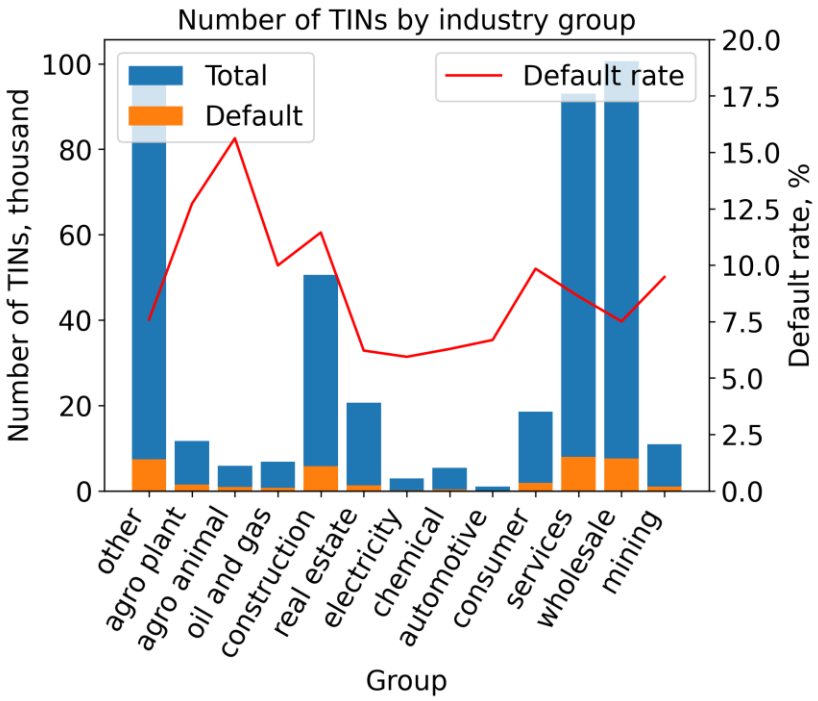


The sectoral affiliation of companies is based on OKVED codes for their main type of activity. The main criteria for assigning a group of OKVED codes to one industry group are the identity of the set of factors affecting the revenue and costs of the companies in the industry, and the similarity of the degree and direction of this influence.

Figure 3 shows the number of companies and the default rate in each industry. There are groups with a small number of companies, such as the electric power industry, the chemical industry, and the automotive industry, and these groups complicate the forecast of the probability of default. There may be inaccuracies in categorisation by OKVED code, such as incorrect accounting of the head offices of companies and their branches. However, this problem is not important for our study, its goal being not to build the best probability of default model as a forecasting metric, nor to take all possible subtleties into account, so we can safely ignore this issue. A more detailed description of the groups can

be found in Table 2 in the Appendix. In the following, we consider the probability of default models for these industry groups.

Figure 3. Number of companies by industry group and default rates in 2012–2018



This classification makes it possible to determine default rates in different industry groups. Consideration of differences among industries usually helps to produce better models. According to this indicator, the most problematic industries in 2012–2018 were the agro-industrial plant growing and animal husbandry sectors. The least risky industries were electricity and real estate.

Bank of Russia Payment System data

The BRPS is a systemically important payment system through which the monetary and budgetary policy of the Russian Federation is implemented. Within the framework of this system, funds are transferred through accounts opened with the Bank of Russia (including those of credit institutions and the Federal Treasury and its territorial bodies).

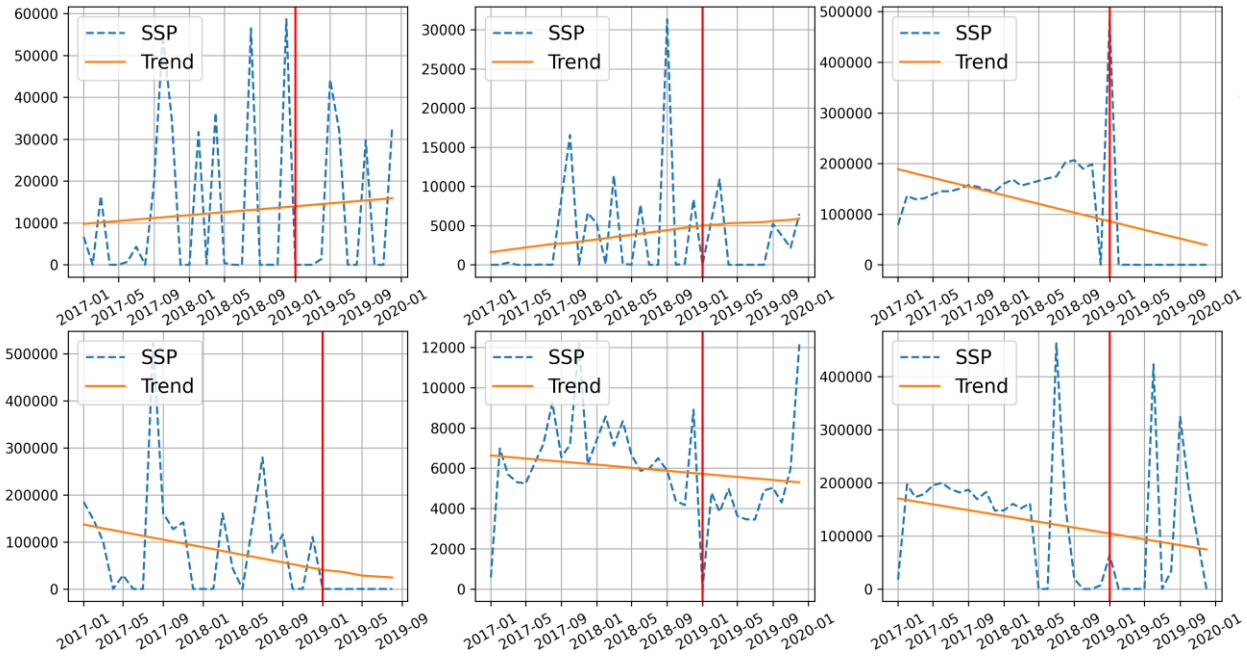
Transaction data for payments can be used as additional information to assess the main indicators of a firm's performance.[4] For example, value-added tax payments can be used as a proxy for the company's revenue, income tax can be a proxy for the organisation’s income, and personal income tax and insurance payments can be a proxy for the payroll.

The main goal of this work is to show the possibility and the usefulness of these data to improve the forecasting quality of standard models for predicting companies’ probability

[4] Transaction data are subject to banking secrecy in accordance with the legislation of the Russian Federation on banks and banking activities.

of default. Using all the available information is not necessary for this purpose. The main idea is to look at aggregated data (mostly tax payments, but other quantities such as total inflows and outflows are also considered as features; see Table 3 in the Appendix for a more detailed description of the indicators). We use BRPS data available to us from 2015 to 2018. An example of aggregated monthly data on payments to the social insurance fund from six different companies from 2017 to 2019 is shown in Figure 4.

Figure 4. Example of BRPS data: aggregated monthly social security payments (SSP) and their trend for six different companies from 2017 to 2018 with a default mark in 2019



At first glance, it would seem that if a company has problems with loan repayment, then it is likely to cut costs, for example, by reducing its staff and therefore its social insurance tax payments. However, the picture above shows that it is not so straightforward. It is difficult to determine from a company's behaviour alone whether it will default or not. In the picture, every company has a default in 2019, but they behave differently. Nevertheless, there may be linear or non-linear relationships in the data that could be useful, and we will use machine-learning methods to mine for useful information.

The BRPS database contains about 10 million transactions per day in 2019, so this can truly be considered ‘big data’, and processing such a large volume of information is indeed a challenge. To avoid computational difficulties, we aggregate transactional data by year and by budget classification code (for features related to budget payments).

To summarise, as baseline features for the prediction of default probability, we use standard features based on accounting data, such as the ratio of working capital to current assets, the ratio of short-term debt to revenue, the receivables turnover ratio, and so on (the full list of 15 variables is given in Table 1 in the Appendix). Our extended feature set

consists of variables from BRPS data (see Table 3 in the Appendix) aggregated by year and normalised by assets.

All values missing from the data are replaced with zeros. Since the indicator can take extremely large positive or negative values for some borrowers (which significantly reduces the quality of the models), the value of each indicator is ranked within its industry. For example, the ratio of debt to profit from sales takes large absolute values if the amount of profit is close to zero. The values of the indicators are then normalised according to their industry group.

4. Methods

To achieve the goal of this paper, our task is to take a method used for the prediction of default probability, apply it to accounting data, and then add variables from the BRPS to the same method and compare the results. In this way, we can prove the usefulness of BRPS data for predicting the probability of default.

Predicting company defaults is a binary classification task: whether the company will default on its loans in the period following the reporting year or not. To be specific, the input data are data for the reporting year, and the target is the mark of default in the year following the reporting year. Therefore, it is necessary to determine the decision threshold, which can lead to difficulties in comparing models. The area under the ROC-curve metric (ROC AUC) is traditionally used for this, since it is free of the problem.

In this section, we describe the machine-learning methods used in this paper to predict the probability of default: the Logistic Regression model (Section 4.1) and the Random Forest model with its feature importance technique (Section 4.2). Section 4.3 contains the details of implementation: the cross-validation scheme, hyperparameter optimisation methods, and methods of working with imbalanced data.

4.1 Logistic Regression

Logistic regression is the most widely used machine learning method in models predicting the probability of default. One of the key advantages of a logistic regression algorithm is that the model can easily be interpreted as a function of the input data. The model consists of coefficients for each variable and an intercept that can be used to explain how the model works.

We use logistic regression with L2 regularisation. The minimisation problem must be solved to find the parameters of the model:

$$\min_{w,c}\left(\frac{1}{2}w^T w + C\sum_{i=1}^{n}\log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right)\right),$$

where $w = (w_0, \dots, w_p)$ is the weights, $X = (x_1, \dots, x_p)$ is the input data, C is the inverse of the regularisation strength, and y_i is the target variable.

To solve the minimisation problem, we use the Limited-Memory BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm described by Byrd et al (1995), enabled in the scikit-learn library for Python.[5]

4.2 Random Forest

Random Forest is a machine learning method consisting of an ensemble of decision tree models proposed by Breiman (2001). Each tree is a tree-structured graph model with nodes as decision points, which are set rules on the explanatory variable for predicting the target variable. The split of decision tree nodes is based on a particular criterion on one of the feature variables, such as Gini (for classification) or sums of squares (for regression), from the entire data set. A leaf node, also called a terminal node, contains a subset of the observations. Splitting continues until a leaf node is formed.

The advantage of the Random Forest method is that it can handle large data sets with higher dimensionality and has high accuracy. This method is more difficult to interpret compared with the Logistic Regression model.

In our paper, we use the implementation of this algorithm from the scikit-learn library for Python.[6] The hyperparameters for optimisation were chosen from among those generally accepted based on the work of Louppe (2014).

Random forests can be used to rank the importance of features in a regression or classification problem. The following technique was described by Breiman (2001) and is implemented in the scikit-learn package in Python. This method is widely used, including for economic problems, for example by Chakraborty and Joseph (2017).[7]

During the training process, the out-of-bag errors for each data point are averaged over the forest. To measure the importance of the *i*-th feature, first, we train the model, then rearrange the values of the *i*-th feature among the training data, and then re-estimate the out-of-bag error. The importance of the *i*-th feature is calculated by averaging the difference in the out-of-bag error before and after permutation across all trees.

4.3 Details of implementation

To obtain robust results and solve the overfitting problem, we use stratified k-fold cross-validation that preserves the ratio of classes in subsets. For testing the model results, we set aside 1/5 of the full dataset (see Figure 5). The training dataset is divided into five subsets, and each of them is used as a validation set. The rest of the training data are used to train the model with the given set of hyperparameters. For each subtest, we evaluate the ROC AUC metric and choose the best model with the given set of

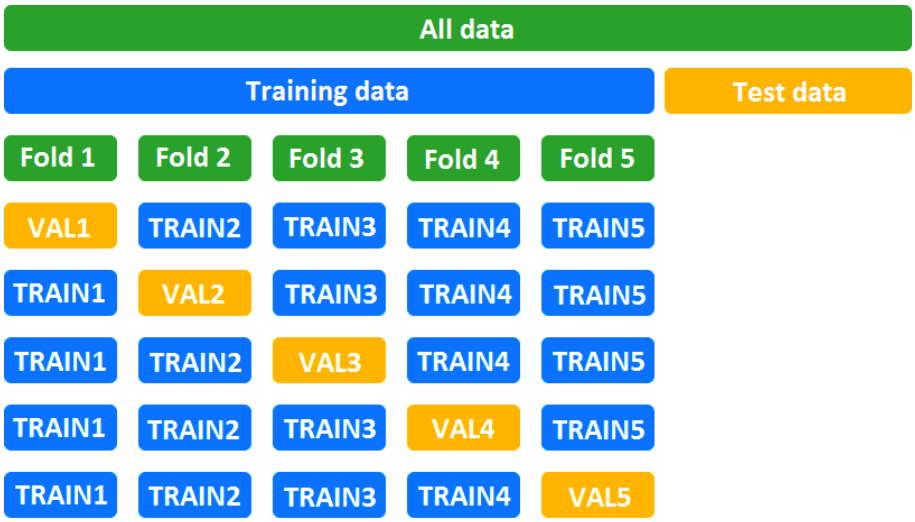
[5] Realisation of Logistic Regression model in the scikit-learn library for Python: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] Realisation of Random Forest Classifier in the scikit-learn library for Python: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] However, impurity-based feature importance can be misleading for features with many unique values. Strobl et al (2007) point out that this technique for computing feature importance is biased. It tends to inflate the importance of continuous or high-cardinality categorical variables.

hyperparameters. To evaluate metrics on the test data, we use all training data with the chosen hyperparameters.

Figure 5. Standard k-fold cross-validation scheme



All comparisons of the model results given in this paper are on the test set according to the cross-validation scheme.

The tuning of the hyperparameters is an important part of the machine learning process. Grid search is one of the most widely used methods for hyperparameter optimisation, but iterating over a large number of parameters takes a lot of time. Bergstra and Bengio (2012) show empirically and theoretically that randomly chosen trials are more efficient for hyperparameter optimisation than trials on a grid.

Dewancker et al. (2016) propose the Bayesian optimisation technique. Bayesian optimisation is a sequential optimisation algorithm that uses the results of the previous iteration to determine the next values of hyperparameters to improve the performance of the model. This approach reduces the number of points and the time it takes to find the best hyperparameters. We use the method implemented in the scikit-learn library for Python.[8]

As described in Section 2, we are dealing with an imbalanced data set. Many machine-learning algorithms rely on the distribution of classes in the training dataset to estimate the probability of observing examples in each class. Such models have therefore poor prediction quality. Techniques designed to change the class distribution by resampling in the training dataset are the most popular solution to the problem of imbalanced classification (Ganganwar (2012); He and Ma (2013); Sonak and Patankar (2015)). The simplest undersampling method is randomly deleting examples from the majority class in the training dataset. The simplest oversampling method randomly duplicates examples in the minority (Figure 6).

[8] Realisation of Bayesian optimiser in the scikit-learn library for Python: https://scikit-optimize.github.io/stable/auto_examples/bayesian-optimization.htm

Figure 6. Example of under- and oversampling technique

Using the undersampling method in our problem leads to a significant reduction in the size of the training sample and an increase in prediction errors. The oversampling method is more reliable for our problem, but it significantly enlarges the dataset by repeating elements and increases the computational time to train the models. In our case, the weighted likelihood function method described by King and Zeng (2001) is the most suitable for logistic regression. This algorithm is implemented in the scikit-learn library.[9] In the Random Forest models, we use the target values to adjust the weights inversely proportionally to the class frequencies in the input data. For more details, see the work of Chen et al. (2003). The algorithm is realised in the scikit-learn library.[10]

5. Results

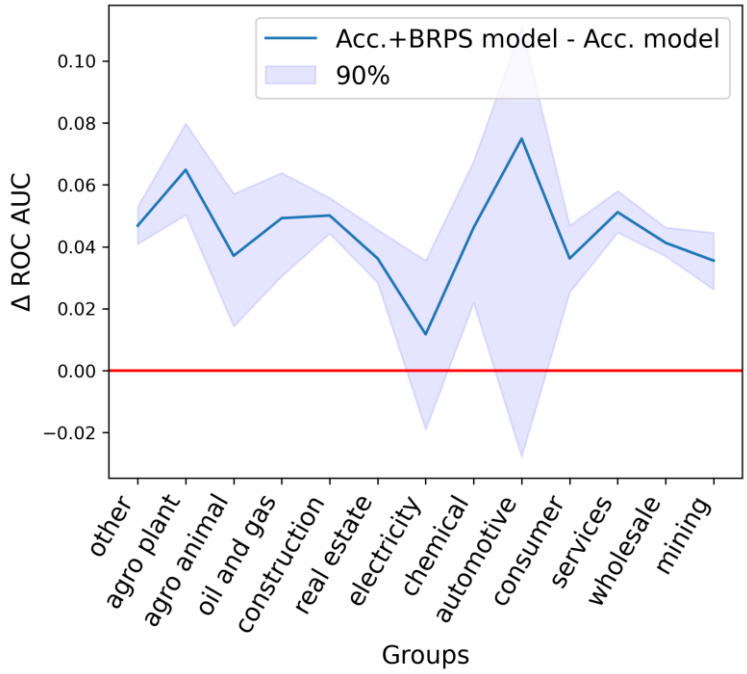
This section summarises the results of the model evaluations. First, we consider the difference in quality according to the ROC AUC metric for forecasts of default probability using Random Forest models with accounting data (the baseline model) and Random Forest models with features extended by the BRPS for each industry group. Section 5.2 presents the results of the comparison of default prediction models based on only the BRPS and the baseline model. Section 5.3 describes the importance of the feature provided by Random Forest models. In Section 5.4, we present the results of a robustness check with the Logistic Regression model with L2 regularisation and the weighted likelihood function. All results are shown for the test data set according to a k-fold cross-validation scheme.

[9] Realisation of Logistic Regression model in the scikit-learn library for Python: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[10] Realisation of Random Forest Classifier model in the scikit-learn library for Python: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

5.1 Random Forest model

To test the usefulness of BRPS, we built Random Forest models using accounting data to predict the probability of default in each industry group. We then built probability of default models for each industry group using the same tools, with the addition of BRPS data. We use the stratified k-fold cross-validation method to optimise the hyperparameters of each model to solve the overfitting problem. As a result, we obtained ROC AUC metrics of 75% and 79% on average across the industry groups for the baseline and extended models, respectively. Figure 7 shows the difference in the ROC AUC metric on the test set of the results of the model for each group, with a 90% confidence interval obtained using the bootstrap method. The results of the regressions in absolute values of the ROC AUC metrics are shown in the Appendix, Figure 1.

Figure 7. Results of the difference in ROC AUC metric on the test set between Random Forest models based on accounting data and Random Forest models based on extended BRPS features

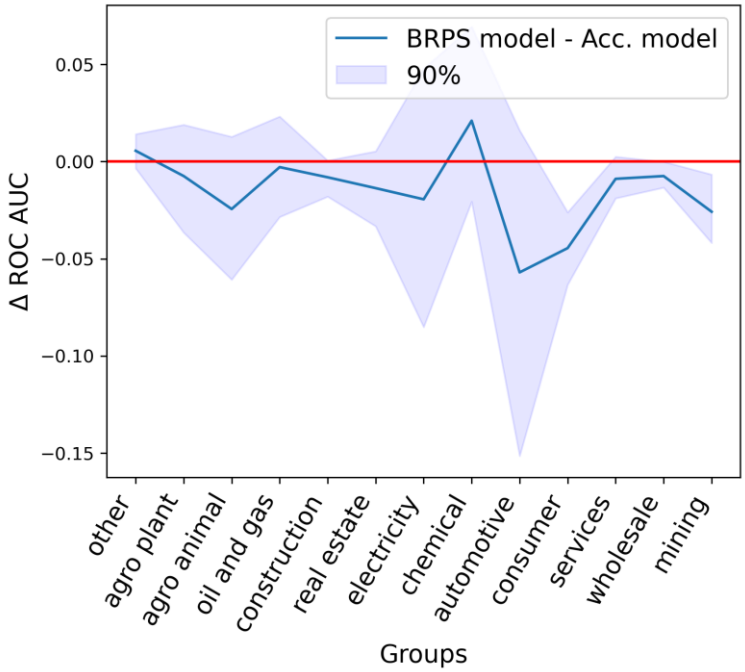


The graph shows that the ROC AUC metrics for the model forecasts with an extended set of variables are higher than the baseline models by 4 points on average for all groups of industries, and for almost all groups the difference is statistically significant. This shows that the BRPS data are useful and informative, and that they can be used to improve the quality of models forecasting Russian companies' probability of default.

5.2 Random Forest model on BRPS data only

In addition, we trained Random Forest models to predict the probability of default in each industrial group using only BRPS data. As a result, we obtained ROC AUC metrics of 75% and 74% on average across the industry groups using the baseline and BRPS data models, respectively. Figure 8 shows the difference in ROC AUC metrics on the test set between the models based on accounting data and the models based on transaction data for each industry group.

Figure 8. Results of difference in ROC AUC metric on the test set between Random Forest models based on accounting data and Random Forest models based on only BRPS data

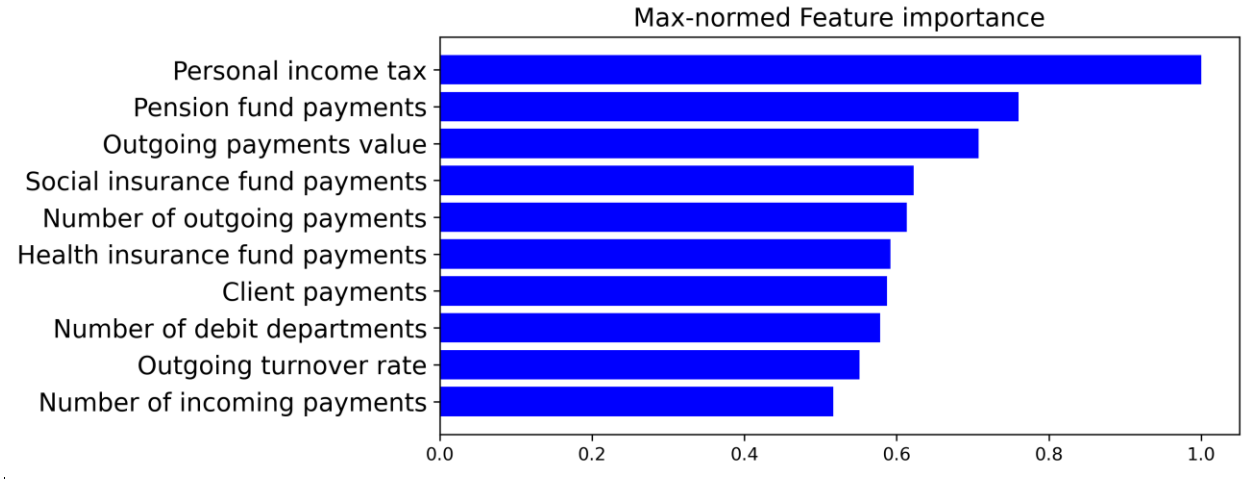


The graph shows that the ROC AUC metric for models based on BRPS data is worse by an average of one point. Nevertheless, these forecast results could be obtained at least three months earlier due to a delay in the publication of annual accounting data. This gives more time for the decision-making process and increases the possibility of the resolution of a situation in which a company has a high probability of default.

5.3 Random Forest feature importance

To assess the importance of BRPS features, we used a Random Forest model trained only on BRPS data. The top-10 max-normed feature importance scores are shown in Figure 9.[11]

Figure 9. Top-10 max-normed feature importance scores of Random Forest classifier

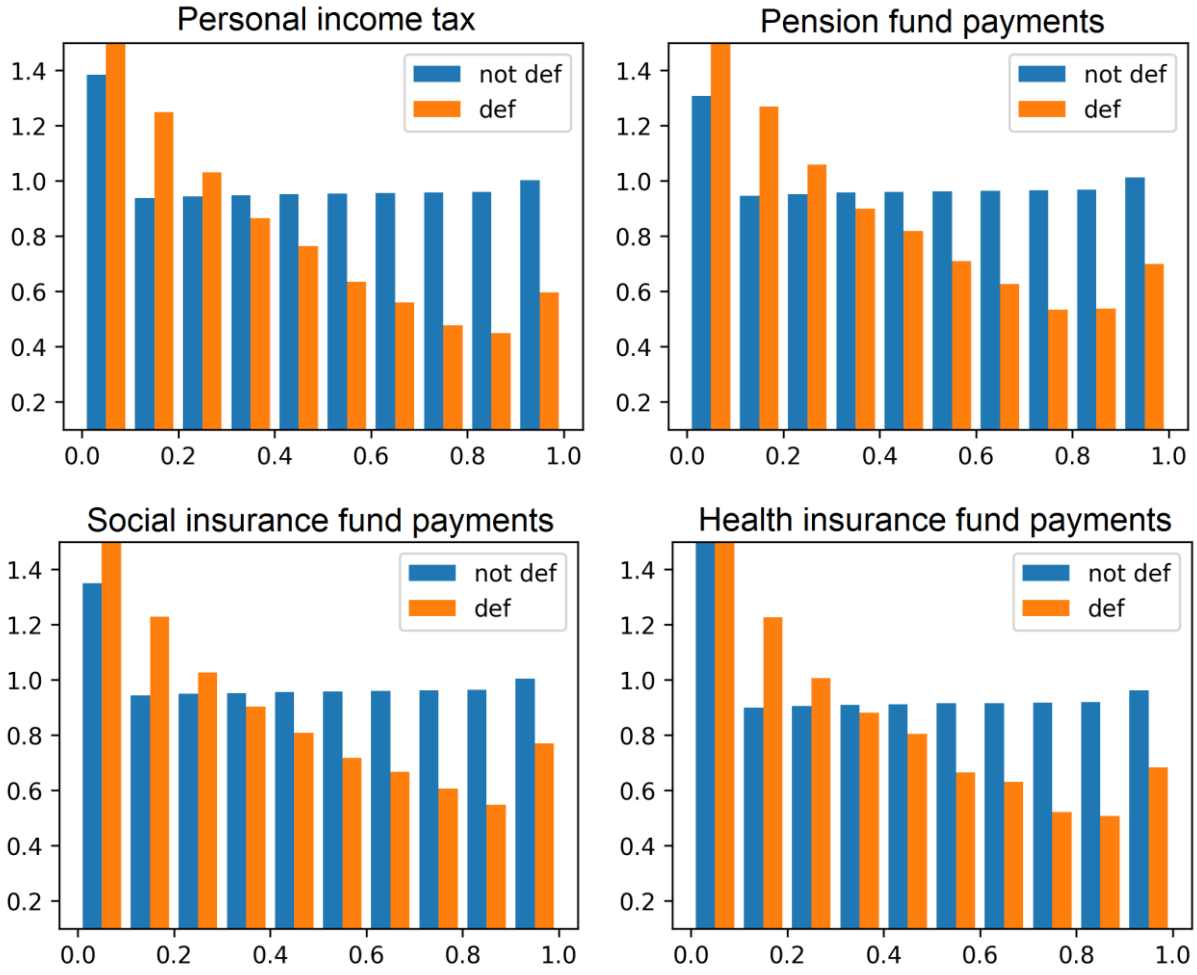


According to these estimates, we can say that the indicators related to payroll (personal income tax payments, pension fund payments, and social insurance fund payments), the value of outgoing payments, and the number of outgoing payments have great importance in predicting the probability of default. This confirms that companies reduce their staffing costs if they begin to have problems with late payments on loans, although not every company does so (see Figure 4).

Figure 4 does not show clear patterns in the dynamics of the salary payments of firms that have problems with loan repayment, but the results of this section reveal that these are the most important features. To see the differences in the important indicators presented above, we built density histograms (see Figure 10).

[11] All accounting variables are in the top 30 according to the Random Forest feature importance algorithm trained on accounting and BRPS data.

Figure 10. Density histograms of personal income tax, pension fund payments, social insurance fund payments, and health insurance fund payments, all normalised by assets



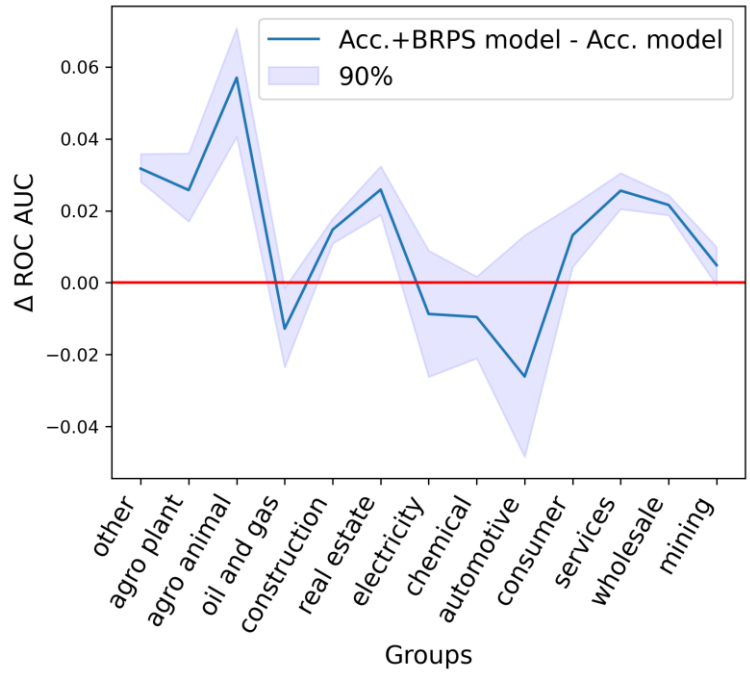
These histograms show that the distributions of these features have different shapes for companies that have a default mark and for those that do not have a default mark in the next year. This difference is precisely the information that is extracted from the data by our models.

5.4 Robustness test with the Logistic Regression model

To check robustness and confirm the results, we also built probability of default models for each industry group using the Logistic Regression method with L2 regularisation and weighted likelihood functions, first using accounting data, and then using the expanded set of features from BRPS data. Since the extended model could not cope with processing the large number of variables even with L2 regularisation, we pre-selected generated features from BRPS data using the Random Forest feature importance technique. We also optimised the inverse of the regularisation strength hyperparameter for each model using the stratified k-fold cross-validation method. As a result, we obtained ROC AUC metrics of 74% and 76% on average across the industry groups for the baseline

and extended models, respectively. Figure 11 shows the difference in the ROC AUC metric of the model results for each group, with a 90% confidence interval obtained using the bootstrap method.

Figure 11. Results of difference in ROC AUC metric on the test set between Logistic Regression models based on accounting data and Logistic Regression model based on extended BRPS features



The graph shows that the ROC AUC metrics for the models with the extended set of features are higher than those of the baseline models almost everywhere, except for the oil, gas, electricity, chemical and automotive industries, where the amount of data in the models is small and the results are highly volatile. The forecast quality of the extended models is 2 points higher according to the ROC AUC metric for all industry groups on average. The results show the importance of BRPS data employed in Random Forest models. The difference in the results of the Random Forest and the Logistic Regression models can be explained by the small number of companies in some groups (and, as a consequence, the higher variance), as well as the lower sensitivity of the Random Forest to outliers in the dataset.

Conclusion

The main goal of this work is to show the usefulness of BRPS data in predicting Russian companies' probability of default. To achieve this goal, we have estimated probability models using machine learning methods and have shown that prediction quality according to the ROC AUC metric is improved by adding BRPS data to accounting data.

Thus, BRPS data contain additional useful information. The Random Forest feature importance showed that the main sources of this additional information are payroll taxes and social payments.

In addition, the use of the BRPS as a separate source of data may also be useful. Although the results of our models trained on BRPS data alone are slightly worse according to the ROC AUC metric than models based on accounting data, BRPS data offer the possibility of obtaining estimates of default probability earlier, which may be especially important in today's fast-changing world.

One direction for future research is to use data from payment transactions with methods based on neural networks, and also assessment on a monthly basis.

References

Altman, E.I. 1968. 'Financial ratios, discriminant analysis and the prediction of corporate bankruptcy'. *The Journal of Finance, 23*(4), 589–609

Babaev, D., Savchenko, M., Tuzhilin, A., and Umerenkov, D. 2019, July. 'E.T.-RNN: Applying Deep Learning to Credit Loan Applications'. In *Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19).* 8 pages. https://doi.org/10.1145/3292500.3330693

Bergstra, J. and Bengio, Y. 2012. 'Random search for hyper-parameter optimization'. *Journal of Machine Learning Research, 13*(2), pp. 281–305

Beaver, W. 1966. 'Financial ratios as predictors of failure'. *Journal of Accounting Research, 4*(3) (Supplement), pp. 71–111.

Beaver, W. 1968. 'Market prices, financial ratios and the prediction of failure'. *Journal of Accounting Research, 6*(2), pp. 179–192.

Breiman, L. 2001 'Random forests'. *Machine learning 45*(1), pp. 5–32. https://doi.org/10.1023/A:1010933404324

Byrd, R.H., Lu, P. Nocedal, J. and Zhu, C. 1995. 'A limited memory algorithm for bound constrained optimization'. *SIAM Journal on Scientific and Statistical Computing, 16*(5), pp. 1190–1208.

Chakraborty, C. and Joseph, A. 2017. *Machine learning at central banks*. Bank of England Working Paper No. 674, 89 pages.

Chen, C., Liaw, A. and Breiman, L. 2003. 'Using random forest to learn imbalanced data'. *University of California, Berkeley, 110*(1–12), p. 24.

Demeshev, B. and Tikhonova, A. 2014. 'Default prediction for Russian companies: Intersectoral comparison'. *HSE Economic Journal, 18*(3), pp. 359—386. (In Russian)

Dewancker, I, McCourt, M. and Clark, S. 2016. Bayesian Optimization for Machine Learning: A Practical Guidebook. https://arxiv.org/abs/1612.04858

Ganganwar, V. 2012. 'An overview of classification algorithms for imbalanced datasets'. *International Journal of Emerging Technology and Advanced Engineering, 2*(4), pp. 42–47.

He, H. and Ma, Y., eds. 2013. *Imbalanced Learning: Foundations, Algorithms, and Applications*. John Wiley & Sons - Technology & Engineering - 216 pages.

Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed*. New York: Springer. 745 pages.

Jackson, R. and Wood, A. 2013. 'The performance of insolvency prediction and credit risk models in the UK: A comparative study'. *The British Accounting Review, 45*(3), pp. 183–202.

Karminsky, A.M. and Burekhin, R.N. 2019. 'Comparative analysis of methods for forecasting bankruptcies of Russian construction companies'. *Business Informatics, 13*(3), pp. 52–66. DOI: 10.17323/1998-0663.2019.3.52.66

King, G. and Zeng, L. 2001. 'Logistic regression in rare events data'. *Political Analysis, 9*(2), pp. 137–163.

Louppe, G. (2014) 'Understanding Random Forests: From Theory to Practice'. PhD Thesis, University of Liege.

Merton, R.C. (1974). 'On the pricing of corporate debt: The risk structure of interest rates'. *The Journal of Finance, 29*(2), pp. 449–470.

Mogilat, A. 2015. 'Bankruptcy in the Russian real sector: Basic tendencies and financial indicators of a typical bankrupt'. *Nauchnye Trudy INP RAN, 13*. pp. 156–186. (In Russian)

Mogilat, A. 2019. 'Modelling financial distress of Russian industrial companies, or What bankruptcy analysis can tell us'. *Voprosy Ekonomiki*, (3), pp, 101–118. (In Russian) https://doi.org/10.32609/0042-8736-2019-3-101-118

Odom, M.D. and Sharda, R. 1990. 'A neural network model for bankruptcy prediction'. In *1990 IJCNN International Joint Conference on Neural Networks* (pp. 163–168). IEEE.

Ohlson, J. A. 1980. 'Financial ratios and the probabilistic prediction of bankruptcy'. *Journal of Accounting Research*, pp. 109–131.

Shibitov, D. and Mamedli M. 2019. 'The finer points of model comparison in machine learning: forecasting based on Russian banks' data'. *Retrieved from Bank of Russia Working Paper Series*, (43). http://www.cbr.ru/content/document/file/87572/wp43_e.pdf

Sonak, A. and Patankar, R. A. 2015. 'A survey on methods to handle imbalance dataset'. *Int. J. Comput. Sci. Mob. Comput., 4*(11), pp. 338–343.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. 2007. 'Bias in random forest variable importance measures: Illustrations, sources and a solution'. *BMC bioinformatics 8(1)*, pp. 1–21. https://doi.org/10.1186/1471-2105-8-25

Appendix

Table 1. Baseline model features

Name	Transcript
K1	Current liquidity ratio
K2	Return on assets
K3	Gross margin
K4	Net profit margin
K5	Operating margin
K6	Equity-to-assets ratio
K7	Debt/earnings from sales
K8	Interest coverage ratio
K9	Interest burden ratio
K10	Borrowed funds/equity
K11	Working capital to current assets ratio
K12	Tax burden ratio
K13	Short-term debt/Revenue
K14	Receivables turnover ratio
K15	Accounts payable turnover ratio

Table 2. List of OKVED codes by industry[12]

Industry name	OKVED codes	Debt as of 1 Jan 2019, billion rubles[13]
Agro-industrial plant growing	01.1	670
Agro-industrial animal husbandry	01.4	1,009
Oil and gas industry	06.1, 06.2, 19.2, 46.71, 46.12.1, 49.50.1, 49.50.2	3,253
Construction	22.23, 25.11, 41, 42, 43	1,501
Real estate operations	68	2,764
Electricity and utilities sector	35	1,302
Chemical industry	20, 21, 22	1,081
Automotive	29	293
Consumer sector	10, 11, 12, 13, 14, 18, 31, 32	1,165
Services sector	45, 47, 49, 50, 51, 52, 53, 58, 61, 62, 63	3,888
Wholesale trade (excluding fuel and minerals)	46	2,128
Metallurgy and mining	05, 07, 08, 09, 23, 24, 46.72, 46.12.2	2,910
Other	Other	5,921
Total		**27,885**

Table 3. Bank of Russia Payment System Data

[12] Borrowers from OKVED 64.20 ‘Activities of holding companies’ were assigned by experts to one of the industries indicated in Table 2.
[13] According to Reporting Form 0409303.

N	Name	Transcript
1	DT	Date of report
2	INN	Taxpayer Identification Numbers
3	OKVED_CODE	Russian Economic Activities Classification System Code
4	CNT_FI_ID_DB	Number of debit departments
5	CNT_FI_ID_CR	Number of credit departments
6	PRC_MAX_DB	Share of the maximum outgoing turnover in percent
7	PRC_MAX_CR	Share of the maximum incoming turnover in percent
8	TOT_DB	Value of outgoing payments
9	TOT_CR	Value of incoming payments
10	CNT_DB	Number of outgoing payments
11	CNT_CR	Number of incoming payments
12	DB06	Land tax
13	DB07	Gambling tax
14	DB08	Property tax
15	DB09	Transport tax
16	DB10	Other taxes
17	DB11	Corporate income tax
18	DB12	Personal income tax
19	DB13	Value-added tax on goods imported to Russia
20	DB14	Value-added tax on goods sold in Russia
21	DB15	Simplified tax
22	DB16	Single tax on imputed income for certain types of activities
23	DB17	Agricultural tax
24	DB18	Tax levied in connection with the patent taxation system
25	DB20	Other payments for social needs
26	DB21	Pension fund payments
27	DB22	Social insurance fund payments
28	DB23	Health insurance fund payments
29	DB24	Federal customs service payments
30	DB30	Other budget payments
31	DB31	Client payments
32	DB32	Payments to non-residents
33	DB33	Settlements with the exchange
34	DB35	Outgoing payments from deposits
35	DB36	Payments to non-resident banks
36	DB37	Foreign currency purchases
37	DB38	Loan repayments
38	DB00	Other outgoing payments
39	CR39	VAT refunds
40	CR40	Payments from the budget
41	CR41	Client payments
42	CR42	Payments from non-residents
43	CR43	Settlements with the exchange
44	CR45	Incoming payments on deposits
45	CR46	Payments from non-resident banks
46	CR47	Foreign currency sales
47	CR48	Receipt of loans
48	CR00	Other incoming payments

Figure 1. ROC AUC metrics of the regressions on the test set with different training features