Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 169 (2020) 767–772

Procedia Computer Science

www.elsevier.com/locate/procedia

Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society)

Technology of forecasting potentially unstable credit organizations based on machine learning methods

Jenny Domashova[a] and Maksim Kulaev[a,*]

[a]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute)
Kashirskoe highway 31, Moscow, 115409, Russian Federation

Abstract

The article presents the results of the application of machine learning methods, in particular, various modifications of decision trees, to predict potentially unstable credit organizations. The application of different modifications of decision trees in the modeling of the specified task and current situation in banking sphere are considered. The technology for solving classification problems using machine learning methods is generalized. A Python program script, which enables to solve classification problems on the basis of the proposed methodology, was developed. The results of the application of machine learning methods using the developed program to solve this problem were described and their quality was analyzed.

© 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures.

Keywords: forecasting; license revocation; machine learning; decision tree

1. Introduction

In 2018, the number of credit institutions operating in Russia decreased by 14% and amounted to 484 as of January 1, 2019, of which 440 were banks. The change in the exchange rate had a significant impact on the dynamics of the

* Corresponding author.
E-mail address: KulaevMA@yandex.ru, janedom@mail.ru

1877-0509 © 2020 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Peer-review under responsibility of the scientific committee of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures.
10.1016/j.procs.2020.02.167

banking sector indicators in 2018 (the nominal effective exchange rate of the ruble against foreign currencies decreased by 8.2% over the year compared to 0.5% in 2017).

Currently, competition in the banking market is at a high level. Credit organizations, despite the impact of the country's largest banks and a significant share of state participation in them, offer their customers additional specialized services. These actions help to attract new consumers, and also gives an opportunity for a bank to find its specific market niche [1].

The main directions of development the financial market report (Bank of Russia, for the period 2016-2018) declared that one of the main tasks is to improve legislation to reduce the number of risks, ensure financial stability of the banking sector and optimize the administrative burden on credit institutions. Moreover, special attention is given to improving the quality of services and raising it to the international level, taking into account the characteristics of the Russian banking sector [2].

Thus, the development of an automated technology for forecasting potentially unstable credit organizations based on machine learning algorithms, which would enable to identify such organization in future, is relevant.

2. Source data

The source data for the study was the Federal Financial Monitoring Service of the Russian Federation information on the activities of commercial banks until 2017. Credit organizations are represented by features that characterize their financial basis, customer activity, suspicious transactions and transactions with fictitious organizations, for instance, percentage of these transactions, the volume of suspicious activity reports (within the framework of AML monitoring, which is crucially important when checking the implementation of AML/CFT law).

First of all, it is necessary to analyze the feature space in terms of the object similarity. In the Russia monopoly banks exist. They take leading positions in the provision of services. The need for these monopolies or the importance of state support for small and regional banks are debatable issues. During an economic crisis, the largest banks with good reputation will receive more customer confidence, as they are more protected from various risks, unlike small regional commercial banks. It is worth noting that the largest players - Sberbank and VTB - differ in size of assets from their closest competitors by several times. Thus, 3 groups of objects can potentially be distinguished: system-forming banks, outsider banks, and all others.

3. Materials and methods

Machine learning methods enable based on the historical training data to customize the algorithm that will identify the main patterns of potentially unstable organizations and will enable them to be identified. There is a vast number of machine learning tasks. As part of the study, the implementation of a supervised learning algorithm is considered to perform a binary classification of new observations.

In particular, the solution of the problem is considered promising based on the "tree" algorithms, which demonstrate the excellent generalizing ability of even the most complex patterns. These include decision tree, random forest, gradient boosting on decision trees, AdaBoost on decision trees.

Hereby, the last three algorithms represent a composition of decision trees, so they can also be attributed to the specified class. Considering a decision tree, it is a set of branches and nodes, where each node splits according to a certain characteristic criterion into several disjoint subsets (in fact, the selection of a subspace in the whole feature space), while all the subsets represent a complete feature space.

According to [3], a random forest is a set of decision trees, the final predictions of which within the framework of solving the considered problems are obtained by a majority vote of each tree.

Considering the gradient boosting based on decision trees it is worth noting that this algorithm consistently build trees so that each next improved the quality of the entire ensemble [4]. This means that at every moment in time the most optimal algorithm is added to the composition, and based on gradient descent and its modifications, the most optimal weights are selected to form the composition.

Within the AdaBoost algorithm [5], in contrast to gradient boosting, each new algorithm in composition focuses on objects that were incorrectly classified by the previous ones, based on their weights, which are inversely proportional to the degree of correctness of the classification of the corresponding objects within the composition.

4. Results

The most common stages, as well as the problems that arise when solving classification problems, are highlighted. The general task of forecasting the potentially unstable credit organizations based on decision trees in the described context is divided into the stages presented in Fig. 1

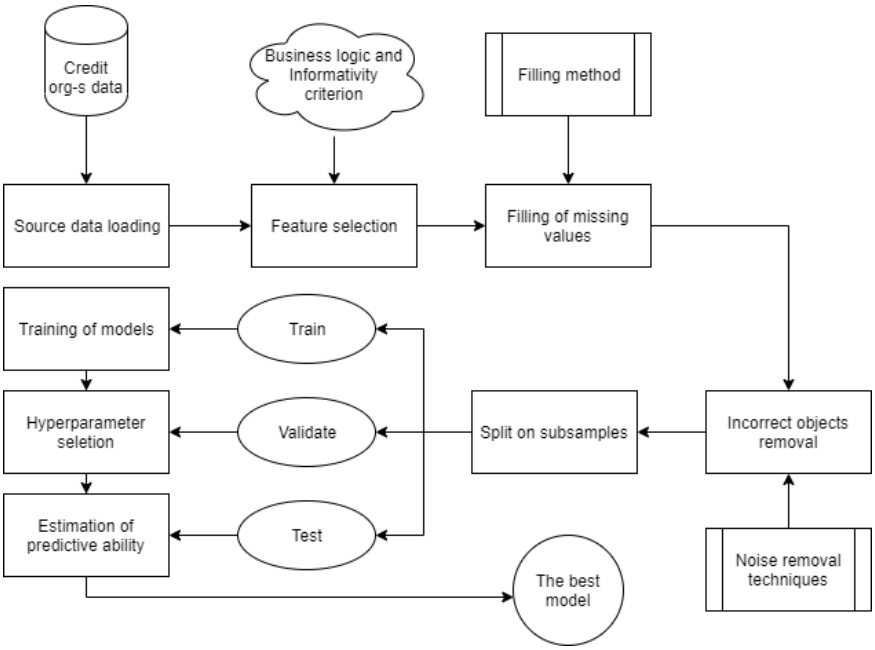

Fig. 1. The research stages

The first stage is not the topic of this study; accordingly, it is believed that this stage was carried out and the input data became available.

The second stage is aimed at identifying the most informative and non-informative features. The former will be the framework for the implementation of the prediction, since they correspond to the greatest information weight when training the model, and the latter, for various reasons, can worsen the final prediction, since they do not provide useful information for the model or depend strongly on more informative features.

At the third stage, the method of filling in the missing values was determined and applied to the source data in order to further implement the prediction on the input object that was not previously known and available.

At the fourth stage, the problem of sampling noisiness was solved by eliminating objects that do not correspond to the considered statistical population using specialized techniques.

The fifth stage involved the specified split of the source data in order to further select the hyperparameters and assess the generalizing ability of the model.

The sixth stage involved learning of the algorithms and then choosing the optimal one.

Feature selection for the analysis was carried out in two steps. Initially, variables were selected based on the business logic of the issue: out of the available groups of features, only those that could provide with useful information to solve the problem were selected. As part of the study, almost all the features were selected, except for the total volumes and the number of operations, because similar information can be obtained from related ones.

Secondly, it should be noticed that as a result of the feature elimination process, no objects with missing values exist. Thus, selected features can actually provide some useful information for the model, and an additional step of filling is not required.

At the next stage, for each interval feature, a search was made for objects whose values were atypical, noise. As part of the study, several methods were studied and applied to search for noise objects, among which the most suitable

for the source data was selected. The identification of atypical objects can be carried out both on the basis of statistical methods of sample analysis and using machine learning methods.

In this paper, we considered methods based on the three-sigma rule of thumb and interquartile distances, statistical tests (Grubbs test, Pierce's criterion, Dixon's Q test), Ellipsoidal Approximation method, low rank matrix approximation, as well as methods using machine learning algorithms, for instance, Support Vector Machine for one class, Isolation Forest, Local Factor method, and cluster analysis [6].

However, it should be noticed that some methods require the normal distribution of data in the sample. When checking this condition by the chi-square criterion at a 5% significance level, the normality hypothesis was rejected, and therefore the following methods were not applied: the Grubbs, Pearson, Dixon methods, Ellipsoidal Approximation.

To test the impact of a particular outlier removal method on the final prediction quality, random forest classifier with automated parameter selection and cross-validation on the existing sample was trained. The analysis of results was performed on the basis of average validation ROC-AUC metric value, a percentage of removed objects and their characteristics. Comparative analysis to determine the best algorithm is presented in table 1.

Table 1. Outliers removal algorithms comparative analysis

Method	Parameters	Validation ROC-AUC value
Without removal	-	0.905
Support Vector Machine	-	0.924
Isolation Forest	Removal percentage = 1%, 5%, 10%, 15%, 20%	0.905 – 0.917
Low rank matrix approximation	-	0.915
3-sigma rule of thumb	-	0.921
Local Factor	Removal percentage = 1%, 5%; number of neighbors = 2, 3, 5, 7, 10	0.896 – 0.905
Cluster analysis	Number of clusters = 2, 3, 4, 5	0.903 – 0.905

Based on the results, it is possible to conclude that the support vector machine for one class demonstrated the best results among advanced outlier removal techniques. The simplest method based on the 3-sigma rule of thumb also showed a decent result. It is also worth noting that some methods give worse results than without removal of outliers. At the same time, Local Factor method, in spite of the lower quality level, singled out the largest market players.

However, for further analysis, we did not use the results of the support vector machine for one class, because it excluded about 600 objects, which was significant for a sample of 1474 objects, and we chose Local Factor method, because it removed those banks, who had a really low probability of license revocation due to the fact that they were major market players. In addition, this method removed only 1% of the objects in the sample, which was comparable with the number of large banks in the Russian Federation and the quality of this method was a small percentage better than just without carrying out the procedure for removing noise objects.

After outlier removal process, three subsamples were finally formed using a stratified random sampling method: training, validation, and test samples. Due to this method, the distribution of the dependent variable in each subset of observations will be proportional to the distribution of the original sample. In addition, the objects into each sub-sample were selected randomly, eliminating negative factors of the use of any patterns. After this split, standardization and normalization (reduction to the standard normal distribution) of all three samples were made, and the average and sample standard deviation for each attribute, calculated on the training sample, was used.

The proposed method for solving the classification problem was implemented using Python programming language.

5. Discussion

An automated selection of the parameters of each model was carried out using cross-validation. Four basic models were considered with the determination of the following hyperparameters:

- Decision tree (maximum depth, minimum number of objects for split);
- Random forest (number of estimators, minimum number of objects for split, maximum tree depth, maximum number of features);
- AdaBoost based on decision trees (maximum tree depth, number of estimators, learning rate);
- Gradient boosting based on decision trees (learning rate, number of estimators, minimum number of objects for split, maximum tree depth, maximum number of features).

The training of the model was carried out on a training set, the selection of hyperparameters was carried out on a validation one. The final decision related to the quality of the model was made on the test data. Evaluation of the quality of training is presented in Tab. 2.

The estimation was made on the basis of the ROC-AUC metric, which is the square under the ROC-curve. Gradient boosting is the best algorithm on both the training and test samples.

Table 2. Estimation of the quality of training models for the training and test samples

Sample	Decision tree	Random forest	AdaBoost	Gradient boosting
Train	0,91	0,96	0,93	0,97
Test	0,89	0,90	0,78	0,92

The ROC-curve on the test sample is represented in Fig. 2.

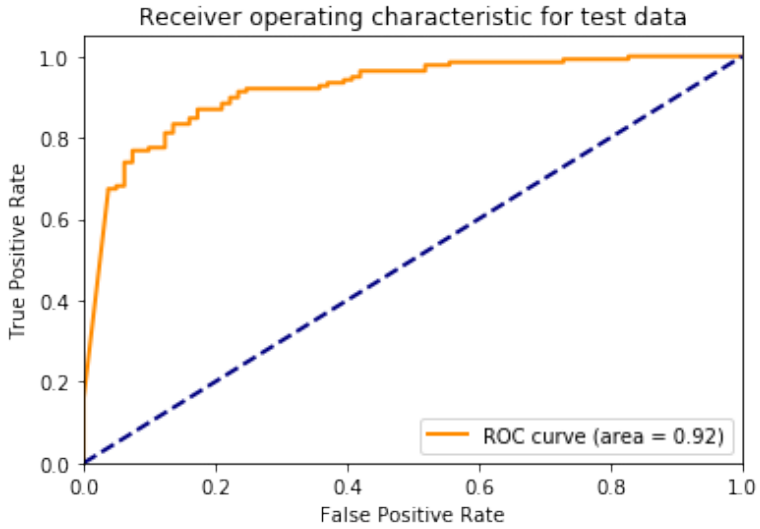

Fig. 2. The ROC-curve on the test sample for the best model

It should be noticed that the value of the area under the learning curve, which demonstrates the predictive ability of the model, turned out to be close to 1. It demonstrates that the model chosen for the final prediction has the high quality.

Final gradient boosting on decision trees model had the following hyperparameters: learning rate 0.01, number of estimators 50, max depth 5, maximum number of features is the squared root of total number of features, minimum number of objects for splitting 2.

6. Conclusion

A method of preparing a training set was proposed and implemented within the framework of solving the problem of forecasting potentially unstable credit organizations. The novelty of the stated approaches to the solution of the considered problem is demonstrated.

A comparative analysis of the predictive ability of various models based on the decision tree and their applicability in general to the solution of this problem is carried out. It was shown that such models are promising for research.

An experimental selection of the best model and optimal parameters has been carried out. The quality of the prepared algorithm was estimated. Moreover, the outlier removal step as a subtask was considered in detail.

The practical significance of the study lies in the possibility of timely identification of credit organizations at risk of revoking licenses, as a result of which measures can be taken in a timely manner to ensure the stability of the functioning of the banking sector of the Russian Federation.

A program of further research will include consideration of other algorithms (both based on the decision tree and not). In addition, other methods for preparing a training set within the framework of business logic will be considered, as well as the use of deep learning models.

References

[1] Bank of Russia (2018). “The development of the Russian Federation banking sector” web resource: https://www.cbr.ru/Collection/Collection/File/14236/razv_bs_18.pdf

[2] Bank of Russia (2016). “The Russian Federation financial market main development directions for the period 2016-2018” web resource: https://www.cbr.ru/finmarkets/files/development/onrfr_2016-18.pdf

[3] Breiman, Leo (2001). Random forests. *Machine Learning*, **45(1)**: 5–32.

[4] Chen, Tianqi and Guestrin, Carlos (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785–794.

[5] Freund, Yoav and Schapire, Robert E. (1997). “A decision-theoretic generalization of on-line learning and an application to boosting,” *Journal of Computer and System Sciences*, **55(1)**: 119 – 139.

[6] Dyakonov, Alexander (2016). “Anomaly detection” web resource: https://dyakonov.org/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/