DOI: 10.17323/j.jcfr.2073-0438.15.3.2021.5-13.
JEL classification: C38, C44, G24, G34

Applying Complementary Credit Scores to Calculate Aggregate Ranking

Zinaida Seleznyova ✉
Lecturer, Research Assistant,
HSE School of Finance, HSE Financial Engineering & Risk Management Lab, National Research University Higher School of Economics, Moscow, Russia,
zseleznyova@hse.ru, ORCID

Abstract

Researchers have been improving credit scoring models for decades, as an increase in the predictive ability of scoring even by a small amount can allow financial institutions to avoid significant losses. Many researchers believe that ensembles of classifiers or aggregated scorings are the most effective. However, ensembles outperform base classifiers by thousandths of a percent on unbalanced samples.

This article proposes an aggregated scoring model. In contrast to previous models, its base classifiers are focused on identifying different types of borrowers. We illustrate the effectiveness of such scoring aggregation on real unbalanced data.

As the effectiveness indicator we use the performance measure of the area under the ROC curve. The DeLong, DeLong and Clarke-Pearson test is used to measure the statistical difference between two or more areas. In addition, we apply a logistic model of defaults (logistic regression) to the data of company financial statements. This model is usually used to identify default borrowers. To obtain a scoring aimed at non-default borrowers, we employ a modified Kemeny median, which was initially developed to rank companies with credit ratings. Both scores are aggregated by logistic regression.

Our data Russian banks that existed or defaulted between July 1, 2010, and July 1, 2015. This sample of banks is highly unbalanced, with a concentration of defaults of about 5%. The aggregation was carried out for banks with several ratings. We show that aggregated classifiers based on different types of information significantly improve the discriminatory power of scoring even on an unbalanced sample. Moreover, the absolute value of this improvement surpasses all the values previously obtained from unbalanced samples.

The aggregated scoring and the approach to its construction can be applied by financial institutions to credit risk assessment and as an auxiliary tool in the decision-making process thanks to the relatively high interpretability of the scores.

Keywords: forward stepwise selection, logistic regression, Kemeny median, credit scoring, statistical learning

For citation: Seleznyova, Z. (2021) "Applying Complementary Credit Scores to Calculate Aggregate Ranking", *Journal of Corporate Finance Research* | ISSN: 2073-0438, 15(3), pp. 5-13. doi: 10.17323/j.jcfr.2073-0438.15.3.2021.5-13.

Introduction

Scoring models have been developing for decades. Researchers have proposed and compared different approaches to data preparation for model construction and approaches to selecting factors which influence credit quality and their generation. They have also studied the best approaches to assessing credit score capability/accuracy and the credit score methods themselves. This was done to improve scoring accuracy, insofar as a gain or loss of a percentage point in accuracy can lead to multimillion profits or losses for banks and other financial institutions [1].

Over the past ten years, scholars have believed that the best practice is to use machine learning models [2] and so-called "ensembles" [3] to construct credit scores. The basic idea of an ensemble lies in the aggregation of modelled base classifiers (scores) with the help of a model/algorithm. There exist different classifications of ensembles [3–5]; however, their division into bagging, boosting, and stacking ensembles is the most common. Bagging is the combination of several independent scorings (base classifiers, weak learners) constructed in a parallel way on the basis of independent random samples. Random forests are a well-known example of bagging. Boosting is the aggregation of several successively constructed base scorings. Stacking is the combination of different base classifiers (for example, logistic regression and a decision tree) that are trained simultaneously. They are combined in the ensemble model (strong learner), which includes different voting rules, statistical models and machine learning methods. The ensemble paradigm makes ensembles relevant: several aggregated classifiers usually show greater discriminatory power/accuracy than a single classifier [5]. Nevertheless, some researchers have shown that ensemble models sometimes fail to surpass machine learning methods in regard to certain criteria [4; 6]. Also, their practical applicability is usually limited: in most cases, they are "black boxes" which are difficult to interpret because machine learning methods and other ensembles are often used as the base classifiers. Therefore, some researchers [7] have attempted to simplify the interpretability of ensemble models, including machine learning methods.

In this paper, we focus on using complementary weak learners to calculate aggregate rankings. We chose the logistic model of defaults and a modified Kemeny median [8] as two weak classifiers of this type due to their relatively high interpretability. We consider them to be complementary for our purposes for the following reasons:

The logistic model (regression) is usually trained for defining default borrowers using corporate financial statements. In other words, the first weak learner is focused on default borrowers.

The modified Kemeny median has been proposed as a tool for credit rating aggregation. Usually, companies which have a better-than-average creditworthiness want to have credit ratings in particular because they are ready to disclose to a rating agency more information than just financial statements. So, this ranking is potentially aimed at non-default companies.

We propose to use logistic regression as the strong learner. It should be noted that logistic regressions, including ridge and lasso, were used as ensemble models in [1; 9] and proved to be superior to other methods considered in these papers.

Our study is based on a sample of banks during the period between July 1, 2010, and July 1, 2015. This sample is characterized by a low default concentration of 5.76%. Financial performance indicators, identifiers of external or government support, and ratings of credit rating agencies were used to create rankings. It was shown that the aggregation of two base classifiers focused on the identification of different types of borrowers results in an improvement of the predictive power of aggregated credit scoring in comparison to base classifiers.

The interpretability of weak and strong learners makes it possible to use aggregate rankings not only as an additional parameter for decision making in financial institutions but also to evaluate default probability in risk management [10; 11]. The proposed weak learners constitute the scientific novelty of this paper: they were trained using potentially complementary information (ratings and financial statements). We know of only one similar study [12] that trained weak learners using market indicators and financial statements. However, the ensemble did not outperform the base classifier in discriminatory power [12].

Literature Review

The number of papers devoted to credit scoring methods has grown exponentially over the past 30 years [3]. In the last five years, researchers have continued their attempts to improve credit scoring for legal entities [13–15]and even more so for financial institutions involved in lending to SMEs. The importance of credit scoring has increased recently because of the financial crisis and increased capital requirements for banks. There are, however, only few studies that develop credit coring models for SME lending. The objective of this study is to introduce a novel, more accurate credit risk estimation approach for SMEs business lending. Based on traditional statistical methods and recent artificial intelligence (AI). However, the majority of papers make use of databases of natural persons [16]. The reason is that such databases are in open access and available for parsing. These samples have been used to compare well-known approaches to credit scoring calculation [17] the volume of databases that financial companies manage is so great that it has become necessary to address this problem, and the solution to this can be found in Big Data techniques applied to massive financial datasets for segmenting risk groups. In this paper, the presence of large datasets is approached through the development of some Monte Carlo experiments using known techniques and algorithms. In addition, a linear mixed model (LMM) and propose new ones [18]. Different ensembles [18; 19] and logistic regressions [20] have been identified as the best scoring methods. In addition, papers dedicated to the comparison of well-

known methods often consider neural networks [21] and decision trees [22] to be the best.

Such a diversity of best methods is partially explained by the wide range of simultaneously applied classification quality criteria. Many authors [4; 9] agree that it is better to use several model performance measures at once. Nevertheless, other researchers [23; 24] continue to apply only conventional methods calculated on the basis of an error matrix.

In this paper, we propose looking at credit scoring aggregation from a slightly different perspective. Usually, only one type of data is used to create base scorings: financial statements or characteristics of natural persons [25]normally taking between 50% and 80% of the total project time. It is in this stage that data in a relational database are transformed for applying a data mining technique. This stage is a complex task that demands from database designers a strong interaction with experts having a broad knowledge about the application domain. Frameworks aiming to systemize this stage have significant limitations when applied to Credit Behavioral Scoring solutions. This paper proposes a framework based on the Model Driven Development approach to systemize the mentioned stage. This work has three main contributions: 1 or company market indicators [13]and even more so for financial institutions involved in lending to SMEs. The importance of credit scoring has increased recently because of the financial crisis and increased capital requirements for banks. There are, however, only few studies that develop credit coring models for SME lending. The objective of this study is to introduce a novel, more accurate credit risk estimation approach for SMEs business lending. Based on traditional statistical methods and recent artificial intelligence (AI. Indicator categories from financial statements complement each other, and machine learning methods can be applied to assess the nonlinear relations between them. However, the creditworthiness of a company may be characterized by factors that are recorded only partially or not at all in statements. These ratings may potentially complement the indicators of corporate financial statements: companies disclose more information to credit rating agencies (CRAs) than one can find in the public domain [26]. In addition, companies with a better creditworthiness, all other things being equal, tend to resort to CRAs: such companies are developing and need external ratings to expand into new markets, for example. Thus, one may conjecture that the collective opinion of credit rating agencies may complement information from financial statements.

In this paper, we will use classical logistic regression as the base classifier and as the aggregated model. This practice was applied in the sample is class imbalanced [9; 27]. Class imbalance may affect the accuracy of default predictions, as classifiers tend to be biased towards the majority class (good borrowers, which showed the advantage of this approach over base classifiers.

In order to calculate base classifiers, a preliminary preparation of data is carried out. One of the stages of preliminary preparation is parameter selection by means of forward feature selection. Nevertheless, it is necessary to describe the data sample before we explain the methodology in detail. This is due to the fact that the choice of methods depends on the data.

Data

The main data pool comprises publicly available information on 958 banks for the period between July 1, 2010, and July 1, 2015, which represents approximately 80% of all banks operating in the Russian Federation during this period. 134 of these banks had two or more ratings calculated by seven credit rating agencies: AK&M, Expert RA (EXP), National Rating Agency (NRA), RusRating (RUS), Fitch Ratings, Moody's Analytics, and Standard & Poor's. This data pool was formally divided into three parts: data on banks up to July 2014, data on banks after July 2014, and data on banks with two and more credit ratings.

Data on banks up to and including July 2014 comprises 70% of the observations of the main pool or 13,570 observations. The default concentration is 4.6%. In terms of default/non-default observations, this sample is highly unbalanced. It comprises indicators from bank report forms 101 and 102 and statutory requirements information (form 135) posted on the website of the Bank of Russia[1] and information on support from the Russian government or foreign banks. This sample was used to train the logistic model of defaults.

Data on banks after July 2014 consists of 4,261 observations with a default concentration of 9.25%. The list of indicators was the same as in the sample described above. This sample was used to test the logistic model of defaults.

Data on banks with two and more credit ratings is part of the two samples described above. This sample consists of observations on 134 banks. The sample size is 1,700 observations, 17 of which are defaults. This sample is also unbalanced and has a default concentration of 2.72%. In addition to the indicators described above, it includes CRA ratings. For the purposes of creating scoring ratings, categories were assigned numerical values, where 0 was attributed to the higher rating category of each credit rating agency (CRA). Then, the numerical value of each lower category was increased by 1. As the last two columns of Table 1 show, the number of assigned rating categories varied greatly from agency to agency.

[1] URL: https://www.cbr.ru/credit/

Table 1. Descriptive statistics of 7 CRA ratings

Variable	Number of observations	Average	Mode	Standard deviation	Min.	Max.
AK&M	209	1.92345	2	0.67502	1	4
EXP	652	1.63497	2	0.87912	0	6
FCH	609	4.92939	0	3.78709	0	14
MDS	1108	6.22563	9	3.12457	0	15
NRA	627	3.70973	3	1.73995	0	13
RUS	246	4.23577	6	2.57644	0	10
SNP	511	5.04305	6	2.8694	0	21

Source: author's calculations.

If we consider previous papers that, in one way or another, studied CRA ratings using Russian data (for example, [28]), we see that the general distribution of agency ratings has changed little. The most frequent ratings are low ratings in the investment grade or best ratings in the speculative grade. The data on ratings is taken from the RUData system[2]. Consensus and aggregate rankings are calculated using this sample.

The low default concentration and small size of the sample of banks with several ratings is insufficient for dividing it into training and test samples to create a logistic model. This is why samples of banks with one or no ratings are used in this study.

Methodology

This chapter consists of several parts. "Logistic Regression" describes the preliminary preparation of data for making a scoring using the logistic model of defaults, the logistic model itself, and ways of validating it. "Modified Kemeny Median" has a similar structure. "Aggregation" describes the mechanism for aggregating the two rankings obtained from the logistic model and the modified Kemeny median. "Model Power Indicator" describes the tool applied to verify the efficiency (power) of obtained rankings.

Logistic Regression

Linear prediction of the logistic model of defaults or the "continuous" rating of the defaults prediction model is used as the first baseline ranking (classifier) [29]. Due to its simplicity, transparency, interpretability and a relatively high discriminatory power, this scoring model continues to be the industry standard [3; 28].

Data preparation. In this paper, observations with missing data were not used for building the logistic regression. Such an approach is frequently used for calculating credit scorings [23; 24], insofar as it does not generate a bias of estimators due to an inappropriately chosen way of imputation of missing values [30]. The forward stepwise selection method was used for features selection for the logistic model. This approach adds a relevant variable to the defined significant variables. If this variable is significant and significantly improves the model, it is also included. In spite of its simplicity, this approach is still widely used to select parameters [16]. Multicollinearity was controlled by means of a correlation matrix. It was controlled both at the intermediate stage of model building and at the final stage.

Logistic regression. In credit scoring problems, the logistic model may be formulated as follows: bank i has rating y_i, which is equal to 0 if there is no bank default and 1 if there is bank default. This rating depends on the latent variable y_i^*, which represents the bank credit quality:

$$y_i = \begin{cases} 1 \ if \geqslant 0 \ (default) \\ 0 \ otherwise (no \ default) \end{cases}, \quad (1)$$

$1 \ if \ y_i^* \geqslant 0$ (default).

$0 \ otherwise$ (no default)

In turn, y_i^* linearly depends on X – factors that may predict the bank creditworthiness. They may be continuous and categorical quantities that represent relevant financial, macroeconomic and other indicators. In this case, the probability of a bank being default or non-default is as follows, respectively:

$$P(y_i = 1) = P(y_i^* \geq 0) = P(X_i'\beta + \varepsilon \geq 0); \quad (2)$$
$$P(y_i = 0) = 1 - F(X_i'\beta),$$

where X_i' is the transposed matrix of factors describing the bank's creditworthiness, ε is an unobservable random component with logistic distribution, and F is a logistic

[2] URL: https://rudata.info/

distribution function. The linear predictions are calculated as follows:

$$R_i^{contin} = \sum_{j=1}^{n} X_i' \beta.$$

In this paper, R^{contin} is used as one of the base scorings focused on default borrowers.

Validation. The complete sample of banks is used to build the logistic model, regardless of whether they have a rating or not. This sample is divided into training and test subsamples. This is done on an out-of-time basis and it's no coincidence. Such a validation method is used in credit scoring studies [31; 32].

Modified Kemeny Median

Data preparation. Unlike the previous method, observations with missing data for certain variables were used for building a modified Kemeny median (consensus ranking). To create a consensus ranking, we used the ratings of seven rating agencies operating in Russia from July 2010 to July 2015.

Modified Kemeny median. Another base classifier is represented by the Kemeny median [8], whose application results in the so-called "consensus ranking". This method is based on the interpretation of credit rating as a relative ranking of objects in accordance with a CRA's opinion on the credit quality of each object. On the basis of the ratings specific nature as expert information, we modified the concept of Kemeni distance between rankings. This made it possible to find a unique solution that least contradicts the opinions of rating agencies with an acceptable accuracy within an acceptable time:

$$R^{cons} = arg\min \sum_{k=1}^{m} \varphi_k \left[\tilde{d}\left(R, R_k\right) + \lambda \delta^2\left(R, R_k\right) \right], \qquad (3)$$

where R^{cons} is the resulting (aggregated) rating, m is the number of aggregated ratings, R_k is the kth rating, $d\left(R', R''\right)$ is the rank measure of distance between ratings R' and R'' (number of contradictory rankings for all pairs of companies), λ is the regularization parameter (relative significance of the secondary criterion), $\delta^2\left(R', R''\right)$ is the additional (secondary) criterion (shows the extent of contradiction significance),

and $\varphi_k > 0, \sum_{k=1}^{m} \varphi_k = 1$

are weights representing the degree of confidence in the ratings of a given agency.

R^{cons} is a non-strict bank ranking. Each combination of ratings is assigned its own numerical value, and so the granularity degree of R^{cons} depends on the number of such combinations, and the order of each combination in R^{cons} depends on its inconsistency with other ratings.

Validation. It is impossible to apply common validation measures such as cross-validation types to this method. The reason is that the modified Kemeny median is a result of a non-parametric approach that cannot be used for another sample directly without mapping.

The Kemeny median was originally a voting method that was subsequently used as an aggregator of credit ratings for banks. The collective opinion of credit rating agencies may complement information from financial statements: companies disclose to credit rating agencies information which may be absent from publicly available data. Thus, it is expected that the combination of the logistic model of defaults built on publicly available data and the ranking obtained from CRA ratings will surpass these base classifiers.

Aggregation

Logistic regression is applied as a strong classifier in this paper. The binary default/non-default variable y_i, arranged in the same way as in function (1), still serves as the interpretable factor. However, to create an aggregated scoring, y_i is predicted using the following two factors:

$$P\left(y_i = 1\right) = F\left(\gamma_0 + \gamma_1 R_i^{contin} + \gamma_2 R_i^{cons} + \varepsilon \geq 0\right), \qquad (4)$$

where γ_j is a coefficient obtained from assessing the logistic regression with the help of the maximum likelihood method and F is the logistic distribution function. The aggregated scoring itself is calculated as follows:

$$R_i^{agregated} = \gamma_0 + \gamma_1 R_i^{contin} + \gamma_2 R_i^{cons}. \qquad (5)$$

Model Power Indicator

We use the indicator of the area under the ROC curve (hereafter, AUCROC) as a measure of the discriminatory power of all scorings. This indicator is appropriate for unbalanced samples – in particular, because it takes different errors into account [1, p. 2]. In addition, this indicator does not underrate or overrate its values due to erroneous classification or default distribution [7, p. 38]. The resulting indicator values should be interpreted as follows: the closer the AUCROC value to 1, the greater the discriminatory power of the credit indicator. This indicator is described in more detail in [33].

The statistical significance of differences between the AUCROC of base classifiers and the aggregated model is defined by means of the DeLong, DeLong and Clarke-Pearson test [34] at a 10% significance level.

Results

This section deals with the discriminatory powers of credit scorings made with the help of base classifiers and through the aggregation of scorings.

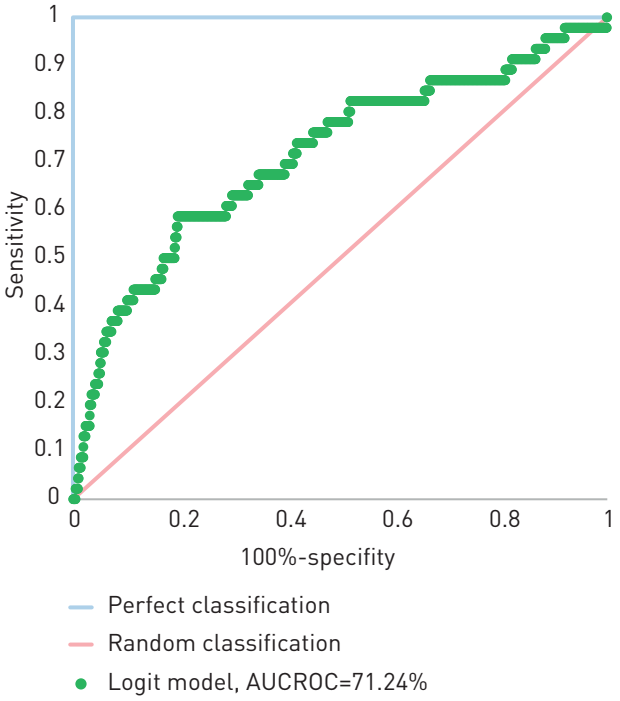
Logistic Model of Defaults

The model was trained on a sample of Russian banks from the period July 1, 2010 – July 1, 2014. The following factors were selected:

1) Ratio of the deposits of a legal entity to its bank assets.
2) Regulatory requirement of "the biggest possible credit risks" H7.
3) Regulatory requirement of long-term liquidity H4.
4) Amount of granted short-term credits.

The AUCROC of the obtained logistic model is equal to 68.26%. In the test sample, the AUCROC is 70.03%. The AUCROC consistency in these two non-overlapping samples indicates that the model has not been retrained. In the sample of banks with two and more ratings, the AUCROC is equal to 71.4% (Figure 1). The quicker growth of ROC diagram at the origin means that the default model defines default banks better.

Figure 1. ROC and AUCROC of the logistic model on a sample of banks with two or more ratings

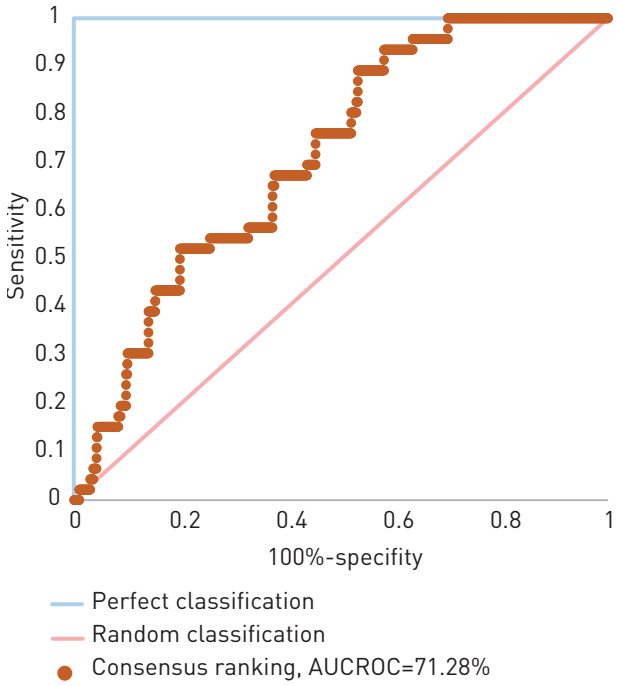


According to the quality criterion of scoring models from [35], this model shows a good discriminatory power from a practical standpoint. This is confirmed by the results of [9; 21], which built logistic regressions using unbalanced samples. In such a case, the AUCROC of the logistic model usually lies in the range 60–74%.

Consensus Ranking

The consensus ranking was calculated on the basis of a sample consisting of banks with two or more ratings. The consensus AUCROC is equal to 71.28% (Figure 2). This ranking defines trustworthy borrowers better, as the right part of the ROC diagram is almost horizontal. The reason for this is that this aggregated rating is based on information about banks which basically have a rating. This is a positive signal for the market: the bank is not afraid of its creditworthiness assessment and can afford it in practice.

Figure 2. ROC and AUCROC of the consensus ranking on a sample of banks with two or more ratings



This ranking also has high discriminatory power from a practical standpoint and is as good as statistical models and machine learning methods in a low-default environment [36; 37]. The consensus ranking is statistically indiscernible at a 10% significance level with a logistic model of defaults according to the DeLong, DeLong and Clarke-Pearson test (p-value = 99.3%).

Aggregate Ranking

The aggregated ranking was built from the two previous rankings. Logistic regression was the aggregated model. We obtained a scoring with the AUCROC equal to 76.16% (Figure 3).

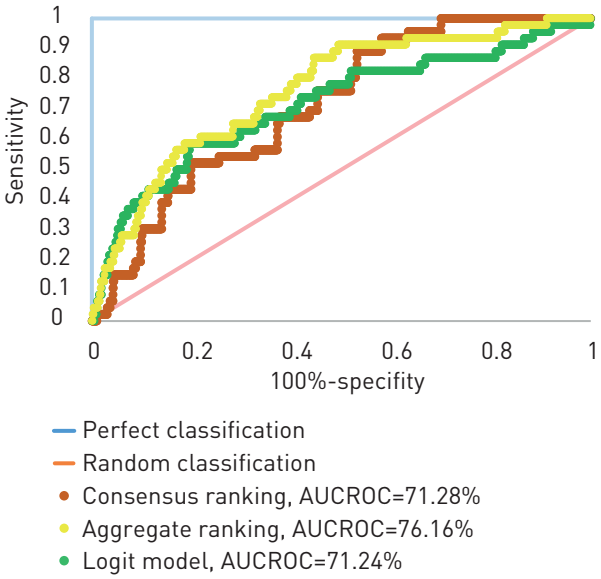
Statistically, the ranking surpasses the two base scorings at a 10% significance level[3], showing the relevance of aggregating several baseline rankings and ensembling. In addition, one should note that aggregated scoring includes the best characteristics of both baseline rankings. It defines default and non-default borrowers with similar precision. Moreover, previously proposed versions of aggregate classifiers showed a growth in the AUCROC not exceeding 3% [38; 39] on unbalanced samples.

[3] $H_0 : AUCROC^{contin} = AUCROC^{aggregated}, p-value = 1.79\%.$

$H_0 : AUCROC^{cons} = AUCROC^{aggregated}, p-value = 9.22\%$.

$H_0 : AUCROC^{contin} = AUCROC^{cons} = AUCROC^{aggregated}, p-value = 0\%$.

Figure 3. ROC and AUCROC of the aggregated model and two base classifiers on a sample of banks with two or more ratings



Conclusion

Financial institutions need to identify both default and non-default contractors or customers in order to enable their management to take informed decisions when solving risk management problems. In this paper, we propose the aggregation of credit scorings made with methods focused on different types of borrowers: the logistic model of defaults and the modified Kemeny median. Logistic regression is used as the strong learner.

Our data sample consists of Russian banks from the period July 2010 – July 2015, including credit ratings. From a practical standpoint, the discriminatory power of baseline rankings is high and typical for credit scorings in a low-default environment. However, their aggregation using logistic regression resulted in a significant growth in the discriminatory power of scoring. Moreover, this increment surpassed the increments of ensembles or aggregated rankings on unbalanced samples described in earlier literature. As long as the applied classifiers demonstrate a relatively high interpretability, such a model can be also used by financial institutions for risk management.

In further research, feature engineering techniques (for example, principle component analysis) may be applied as explanatory factors, provided the obtained index is interpretable. It is also possible to expand the set of base scorings by adding market scorings and some other interpretable scorings obtained, for example, from discriminant analysis, decision trees, etc.

References

1. Abellán J, Castellano JG. A comparative study on base classifiers in ensemble methods for credit scoring. Expert Syst Appl. 2017 May 1;73:1–10. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417416306947
2. de Castro Vieira JR, Barboza F, Sobreiro VA, Kimura H. Machine learning models for credit analysis improvements: Predicting low-income families' default. Appl Soft Comput J. 2019 Oct 1;83:105640.
3. Louzada F, Ara A, Fernandes GB. Classification methods applied to credit scoring: Systematic review and overall comparison. Vol. 21, Surveys in Operations Research and Management Science. Elsevier Science B.V.; 2016. p. 117–34.
4. Xia Y, Liu C, Li YY, Liu N. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst Appl. 2017 Jul 15;78:225–41.
5. Ensemble methods: bagging, boosting and stacking | by Joseph Rocca | Towards Data Science [Internet]. [cited 2020 Oct 4]. Available from: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205
6. Ma X, Sha J, Wang D, Yu Y, Yang Q, Niu X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron Commer Res Appl. 2018 Sep 1;31:24–39. Available from: https://linkinghub.elsevier.com/retrieve/pii/S156742231830070X
7. Hayashi Y. Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. Oper Res Perspect. 2016 Jan 1;3:32–42.
8. Buzdalin AV, Zanochkin AU, Kurbangaleev MZ, Smirnov SN. Credit ratings aggregation as a task of building consensus in the system of expert assessments. Globalnie rinki i finansoviy engineering. 2017;4(3):181–207. (In Russ.).
9. Zanin L. Combining multiple probability predictions in the presence of class imbalance to discriminate between potential bad and good borrowers in the peer-to-peer lending market. J Behav Exp Financ. 2020 Mar 1;25:100272.
10. Committee on Banking Supervision B. Basel Committee on Banking Supervision International Convergence of Capital Measurement and Capital Standards A Revised Framework Comprehensive Version [Internet]. 2006 [cited 2020 Oct 4]. Available from: https://www.bis.org/basel_framework/
11. Carta S, Ferreira A, Reforgiato Recupero D, Saia M, Saia R. A combined entropy-based approach for a proactive credit scoring. Eng Appl Artif Intell. 2020 Jan 1;87:103292.
12. Tserng HP, Chen PC, Huang WH, Lei MC, Tran QH. Prediction of default probability for construction firms using the logit model. J Civ Eng Manag. 2014;20(2):247–55.
13. Li K, Niskanen J, Kolehmainen M, Niskanen M. Financial innovation: Credit default hybrid model for SME lending. Expert Syst Appl. 2016;61.

14. Barboza F, Kimura H, Altman E. Machine learning models and bankruptcy prediction. Expert Syst Appl. 2017 Oct 15;83:405–17.
15. Maldonado S, Peters G, Weber R. Credit scoring using three-way decisions with probabilistic rough sets. Inf Sci (Ny). 2020 Jan 1;507:700–14.
16. Dastile X, Celik T, Potsane M. Statistical and machine learning models in credit scoring: A systematic literature survey. Appl Soft Comput J. 2020 Jun 1;91:106263.
17. Pérez-Martín A, Pérez-Torregrosa A, Vaca M. Big Data techniques to measure credit banking risk in home equity loans. J Bus Res. 2018 Aug 1;89:448–54.
18. Ala'raj M, Abbod MF. A new hybrid ensemble credit scoring model based on classifiers consensus system approach. Expert Syst Appl. 2016;64.
19. Ala'Raj M, Abbod MF. Classifiers consensus system approach for credit scoring. Knowledge-Based Syst. 2016 Jul 15;104:89–105.
20. Fang F, Chen Y. A new approach for credit scoring by directly maximizing the Kolmogorov–Smirnov statistic. Comput Stat Data Anal. 2019 May 1;133:180–94.
21. Teply P, Polena M. Best classification algorithms in peer-to-peer lending. North Am J Econ Financ. 2020 Jan 1;51:100904.
22. Butaru F, Chen Q, Clark B, Das S, Lo AW, Siddique A. Risk and risk management in the credit card industry. J Bank Financ. 2016 Nov 1;72:218–39.
23. Sousa MR, Gama J, Brandão E. A new dynamic modeling framework for credit risk assessment. Expert Syst Appl. 2016 Mar 1;45:341–51.
24. Maldonado S, Pérez J, Bravo C. Cost-based feature selection for Support Vector Machines: An application in credit scoring. Eur J Oper Res. 2017 Sep 1;261(2):656–65.
25. Neto R, Jorge Adeodato P, Carolina Salgado A. A framework for data transformation in Credit Behavioral Scoring applications based on Model Driven Development. Expert Syst Appl. 2017 Apr 15;72:293–305.
26. Karminsky A, Polozov A. Handbook of Ratings: Approaches to Ratings in the Economy, Sports, and Society. Handbook of Ratings: Approaches to Ratings in the Economy, Sports, and Society. Springer International Publishing; 2016. 1–356 p.
27. 27. Li K, Niskanen J, Kolehmainen M, Niskanen M. Financial innovation: Credit default hybrid model for SME lending. Expert Syst Appl. 2016 Nov 1;61:343–55.
28. Karminsky A.M. Credit rating and its modelling. Moscow: HSE Publishing House; 2015. 304 p. (In Russ.).
29. Aivazyan S, Golovan S, Karminsky A, Peresetckiy A. Approaches to comparing rating scales. Prikladnaya ekonometrika. 2011;3(23):13–40. (In Russ.).
30. Fitzmaurice G, Kenward MG. Handbooks of Modern Statistical Methods Handbook of Missing Data Methodology. 1st ed. New York: Chapman and Hall/CRC; 2014. 600 p.
31. Mushava J, Murray M. An experimental comparison of classification techniques in debt recoveries scoring: Evidence from South Africa's unsecured lending market. Expert Syst Appl. 2018 Nov 30;111:35–50.
32. Garrido F, Verbeke W, Bravo C. A Robust profit measure for binary classification model evaluation. Expert Syst Appl. 2018 Feb 1;92:154–60.
33. Tasche D. Estimating discriminatory power and PD curves when the number of defaults is small. 2009 May 24 [cited 2018 Dec 5]; Available from: http://arxiv.org/abs/0905.3928
34. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988 Sep;44(3):837.
35. Pomazanov M.V. Credit risk-management in a bank: internal ratings-based approach (IRB) [Internet]. 1st ed. Penikas G.I, editor. Moscow: Urait; 2019 [cited 2020 Oct 4]. 265 p. Available from: https://urait.ru/book/upravlenie-kreditnym-riskom-v-banke-podhod-vnutrennih-reytingov-pvr-437044. (In Russ.).
36. Fitzpatrick T, Mues C. An empirical comparison of classification algorithms for mortgage default prediction: Evidence from a distressed mortgage market. In: European Journal of Operational Research. Elsevier; 2016. p. 427–39.
37. Shen F, Zhao X, Li Z, Li K, Meng Z. A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation. Phys A Stat Mech its Appl. 2019 Jul 15;526:121073.
38. Xia Y, Liu C, Da B, Xie F. A novel heterogeneous ensemble credit scoring model based on bstacking approach. Expert Syst Appl. 2018 Mar 1;93:182–99.
39. He H, Zhang W, Zhang S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Syst Appl. 2018 May 15;98:105–17.

Contribution of the authors: the authors contributed equally to this article.
The authors declare no conflicts of interests.

The article **was submitted** 06.07.2021; **approved after reviewing** 08.08.2021; **accepted for publication** 14.08.2021.