Research in International Business and Finance 70 (2024) 102397

Contents lists available at ScienceDirect

Research in International Business and Finance

journal homepage: www.elsevier.com/locate/ribaf

Credit scoring: Does XGboost outperform logistic regression?A test on Italian SMEs

Stefano Zedda

University of Cagliari, Department of Economics and Business Sciences, Italy

ARTICLE INFO

Keywords:
Credit scoring
Logistic regression
XGBoost
Bank lending
SMEs

ABSTRACT

The old-fashioned logistic regression is still the most used method for credit scoring. Recent developments have evolved new instruments coming from the machine learning approach, including random forests.

In this paper, we tested the efficiency of logistic regression and XGBoost methods for default forecasting on a sample of 35,535 cases from 7 different business sectors of Italian SMEs, on a set of 28 banking variables and 55 balance sheet ratios for verifying which approach is better supporting the lending decisions.

With this aim, we developed an efficiency index for measuring each model's capability to correctly select good borrowers, balancing the different effects of refusing the loan to a good customer and lending to a defaulter. Also, we computed the balancing spread to quantify the different models' efficiency in terms of credit costs for the borrower firms.

Results show that different sectors report different results. However, generally speaking, the two methods report similar capabilities, while the cutoff setting can make a substantial difference in the actual use of those models for lending decisions.

1. Introduction

Bank lending is based on a repeated evaluation activity, in which the possible gain from interests is (less than) balanced by the losses due to defaults. Thus, the choice of whether to lend money or not to a firm and, more precisely, measuring the risk associated with any loan and managing the associated credit risk is the key banking activity.

Credit risk can be defined as the risk that the borrower does not timely meet its debt obligations. Credit risk measurement and management has been one of the key points since the first Basel Committee capital accords of 1988 in which credit risk was proxied by categories. However, it became the core of regulatory provisions from the Basel II accords of 2004. Since then, the minimum capital requirements for banks are computed based on their risk-weighted assets (RWA), so on the specific risk weighting of each loan, to be estimated by the bank.

In these models, the main estimation problem refers to the loan's probability to default (PD), possibly integrated with the loss-given-default (LGD) estimation.

For modeling the PD estimation, previous data on each loan and borrower characteristics are coupled with its default/non-default results. The loan and borrower firms' characteristics typically include balance sheet data, credit lines and current account data, and credit records about the previous and current loans.

E-mail address: szedda@unica.it.

https://doi.org/10.1016/j.ribaf.2024.102397
Received 31 December 2023; Received in revised form 7 March 2024; Accepted 14 May 2024
Available online 24 May 2024

According to Llewellyn (2012), *Bank business models are not static and evolve over time and under the influence of a complex mix of exogenous and endogenous pressures. The more powerful of these pressures include: the structural evolution of the financial system and financial markets; the macro-economic environment in which banks and their customers operate; regulation; the competitive environment in banking markets; financial innovation, and the chosen business objectives of banks (e.g. asset growth, market share, ROE, etc.)*.

Among the signaled pressures, the recent evolution of machine learning models is one of the technological innovation opportunities that can bring more efficiency and speed to the banks' lending decisions, evolving their business models from a human-centered evaluation to a more extensive use of a quantitative, algorithms-based evaluation, as it already happened for consumer credit.

While this risk assessment has been the main activity done by bank experts for some centuries, since the pioneer works of Durand (1941) and Beaver (1966) on the use of statistical methods for this purpose and the seminal paper of Altman (1968), quantitative credit scoring has been developed in different ways and reported significant improvement. Therefore, quantitative models have progressively substituted the traditional human evaluation of this key variable, which can provide a timely, stable, monitorable, and cheap decision method (Blochlinger and Leippold, 2006).

After the initial use of multivariate linear discriminant regression, the logistic discriminant regression, due to its theoretical coherence to the problem of estimating a dichotomic variable, gained the stage as the reference method for default forecasting.

Credit scoring models typically assess the creditworthiness of loans through their PD estimate, based on the characteristics of the loan and the borrower, possibly integrated by data about the business cycle and the economic framework.

Logistic regressions can determine the linear relationship between the default occurrences and the considered variables but cannot catch the nonlinear relationship between the dependent and independent variables (Banasik et al., 2003).

Stylized facts of the literature in this stream are the need for a contextualized estimation and tuning, as to say, the specific estimation for each country, firms' activity sector, and dimension resulted in being of fundamental importance for qualified model accuracy, e.g., Rikkers and Thibeault (2011) report that prediction models developed for specific activity sectors perform better than generic models and that the specific estimation of models on SME samples increases the default prediction accuracy rates for this kind of firms (Altman et al., 2020; Gupta et al., 2015).

Other recent papers that have used methodologies from the logit family to study SME finance issues, e.g., Mascia and Rossi (2017), Galli et al. (2018) and Galli et al. (2020) tested for gender inequality; Mascia (2018) verified that new enterprises experience higher denial rates from banks compared to more established businesses; Filipe et al. (2016), in their interesting attempt to integrate accounting-based and macro-economic variables for SME default prediction, highlight the importance of developing the scoring models on a regional basis.

Over time, these methods were compared with the statistical research outcomes and artificial intelligence methods, including machine learning, neural networks, support vector machines, and random forests, which can capture the nonlinear relationship among variables.

A review of the different approaches and results in this stream are in Baesens and Van Gestel (2009) and Doumpos, Lemonakis, Niklis, and Zopounidis (2019).

Still, logistic regressions are widely used due to their simple application and comprehensibility. Even large datasets can be easily treated by these techniques, and their results, when just considering the linear effects, are straightforward to understand and transform into a model/algorithm to be used in practice for actual banking activity. The limits of this approach are its incapability to identify complex risk patterns involving more variables in a nonlinear way, which, instead, can be identified by more complex methods, whose complexity typically results in higher performance but incomprehensible (black boxes) models. This fact is a major issue not only from the scientific point of view, which is always aimed at understanding the subjacent process, but also from the supervision and regulation point of view.

A key point in this debate is the comparative performance of the different methods, as a significantly higher efficiency can overcome the black box limits.

Within machine learning methods, Chen and Guestrin's (2016) XGBoost ensemble learning gained strong attention due to its excellent performances, and recently applied to financial issues, for fraud detecting (Nti et al., 2022; Liu, 2023; Hajek et al., 2023), credit risk assessment (Marques et al., 2013; Lessmann et al., 2015; ZHANG et al., 2023; ZHANG et al., 2023; Xueping et al., 2023), and other related issues (Zhiyao et al., 2024; Maehara et al., 2024).

Instead of using only one model, Ensemble methods are based on the estimate of multiple models. The method Breiman (2001) suggested, known as Random forests, is based on a set of decision trees (thus, a forest) created within the training phase. In the testing phase, the forest outputs the result predicted by the majority of trees. For a significant distinction of the decision trees, each tree sees just a random sample of the training set and a random selection of the input variables. In this way, the constructed decision trees are based on different evaluations, and the performance of the ensemble of decision trees results in being more accurate than any individual base model.

Within this family, the XGBoost is based on an ensemble of decision trees, optimized based on a loss function, for measuring how well the model fits the data to reduce the loss function to its lowest values. Just a few papers used this method for credit scoring, among them, a significant contribution came from He, Zhang, and Zhang (2018) and Xia et al. (2017).

In this paper, we performed logistic regression and XGBoost credit scoring estimations on a sample of Italian SMEs to compare their selection performances and verify if the newer and more complex method can actually overcome the older one.

To compare the two, we adopted a machine learning approach, training the models on a part of the data and testing the two models' performances out-of-sample on the residual part of the data.

With this aim, we developed an efficiency index for measuring each model's capability to correctly select good borrowers, balancing the different effects of refusing the loan to a good customer and lending to a defaulter. The reference scale is set to have the maximum reference (100 % efficiency) to the theoretically optimal model, as to say, the one correctly selecting all customers, and the minimum reference (0 % efficiency) for the theoretically worst model, failing all evaluations.

Also, we computed the balancing spread to quantify the different models' efficiency in terms of credit costs for the borrower firms.

Results show that the two methods report remarkably similar results. At the same time, other sides of the estimation, such as selecting a different cut point, can be more significant for actual banking use.

The remainder of this paper is structured as follows: Section 2 describes the considered data sample and variables, the data cleaning process, and the methodology used for the estimations; Section 3 reports its results, both in terms of direct evaluation of forecast capabilities and in terms of its balancing and optimization for a possible banking use. Section 4 discusses the significance of the results and concludes.

2. Methodology and data

2.1. Data

As anticipated in the introduction, the analysis is developed on a sample of 35,535 Italian SMEs, selected from a unique dataset of more than 74.000 balance sheets of Italian firms coming from 113 local banks from 2011 to 2014, exclusively available from the IT services company belonging to one of the two Italian cooperative banking groups.

One of the crucial characteristics of this dataset is that it includes internal banking data and that default events are recorded in case of unregular fulfillment of bank debts, allowing for specific detection of credit defaults and assessing the role of banking relationship data in default forecasting.

These data are typically highly confidential as they are crucial in assessing the firm's creditworthiness and determining a competitive advantage over competitor banks. In this case, the loss of competitive advantage arising from the public availability of results was limited by the data source just releasing the data referring to the three years from 2011 to 2014.

Concerning the sample significance, the diversification of Italian regions' economic structure, the per capita GDP of 2014 ranging from €16,200 to €39,900, makes this sample an interesting proxy for the whole of Europe, including in the same value range as the average per capita GDP of 14 European countries for the same year.[1]

For our estimation, we considered the business sectors more represented in our dataset, and performed both a logistic stepwise regression and an XGBoost estimation.

The considered business sectors are classified by the Italian classification system (Ateco code) as 25 “Production of metal products”, 28 “Production of machinery and equipment", 41 “Construction of buildings”, 43 “Specialized construction works”, 45 “Wholesale and retail of cars and motorcycles”, 46 “Wholesale, excluding cars and motorcycles”, 47 “Retail, excluding cars and motorcycles”.

For each case, our dataset included both banking data and financial ratios, reporting a complete view of the firm economic results, financial equilibriums, and banking relationships.

The selection of cases from SMEs of one specific activity sector gives a higher significance to the credit scoring outcome because, as reported in the introduction, it refers to dimensionally homogeneous cases and firms with similar activities characterized by similar balance sheet equilibriums. For a proper analysis of the actual banking use of these methodologies, the default is set as the case of an unregular refund of bank loans (past due 90), and not to the firm liquidation, which, in case, occurs significantly later and is not necessarily related to the bank-firm relationship. The sample (see Table 1) includes a significant number of defaults for each of the considered business sector, so nicely suited for our estimations.

The dataset includes both balance sheet variables and banking variables. The 55 balance sheet variables (see Table A1 in the Appendix) include economic, financial, and structural ratios to consider different aspects of the firm balance sheet. The 28 banking variables monitor various credit relationships aspects, such as credit availability, the actual use of credit, blank cheques, overdrafts, and credit and debit credit account movements.

2.2. Data cleaning

To analyze the available data, we first verified which variables reported unavailable data and possible outliers to select the most qualified variables. As shown in Table 3, variables can include unavailable values (NA's) for some cases and extreme values, which would depreciably influence the estimations. The logistic regression can only consider the cases when all variable values are available, so any cell that is not available excludes the whole data line. As different availabilities are reported on different variables, the occurrence of, say, 10 % of values for each variable can seriously limit the applicability of this method already to a set of 10 variables.

We thus performed a data cleaning, starting from the distribution of values exemplified in Table 2.

After excluding the unacceptable values, we performed a winsorization, so to bring the outliers above the 99 % or under the 1 % probability distribution to these thresholds. In this way, the large (or small) value on the variable of extreme observations is partially preserved, but its deviating influence on the estimations is limited. Comparing Table 2 to Table 3, we can see, e.g. in the case of “Fixed asset coverage”, that the maximum value is reduced from the extreme value of 6,565,258, to a more reasonable, but still very high value of around 146,302. This process preserves the core distribution, so the values from the 1st quarter to the 3rd quarter are not

[1] The countries are Austria, Belgium, Cyprus, Finland, France, Germany, Greece, Italy, Malta, Netherlands, Portugal, Slovenia, Spain, and the UK, source: Eurostat.

Table 1
Data description.

Business sectors	No default	Default	Total
25 - Production of metal products	4,800	219	5,019
28 - Production of machinery and equipment	2,270	99	2,369
41 - Construction of buildings	5,288	792	6,080
43 - Specialized construction works	4,097	317	4,414
45 - Wholesale and retail of cars and motorcycles	2,359	114	2,473
46 – Wholesale (excluding cars and motorcycles)	10,025	516	10,541
47 – Retail (excluding cars and motorcycles)	4,340	299	4,639
Total	**33,179**	**2,356**	**35,535**

Note: All considered sectors include more than 2,000 cases, and a significant share of defaults.

Table 2
Some variables distribution summary before cleaning.

	Added value on production value	Depreciation and devaluation on costs	Payables to banks on current assets	Fixed asset coverage
Min	0	0	0	0
1st Quartile	22.2	0.7	30.8	46.9
Median	40.9	1.7	56.7	153.0
Mean	46.5	5.7	1330.1	11,703.0
3rd Quartile	70.6	4.3	84.4	863.5
Max	135.0	1,023.4	1,605,126.0	6,565,258.0
NA's	680	1073	879	796

Note: Before the cleaning, all variables include a significant number of "NA's", and often, the mean value is higher than the 3rd quartile due to the presence of outliers, as signaled by huge Max values.

Table 3
Variables distribution after winsorization.

	Added value on production value	Depreciation and devaluation on costs	Payables to banks on current assets	Fixed asset coverage
Min	1.4	0.02	0.6	1.7
1st Quartile	22.2	0.7	30.8	46.9
Median	40.9	1.7	56.7	153.0
Mean	46.5	5.4	182.4	4316.1
3rd Quartile	70.6	4.3	84.4	863.5
Max	100,0	65.6	6,965.9	146,302.4
NA's	680	1073	879	796

Note: After winsorization, the Max values are significantly reduced, and subsequently also the mean values are reduced.

influenced, and the mean value is typically brought nearer to the median value.

Then, to maximize the available information without biasing the results, we replaced the not available values with the mean value. This process does not influence the mean value. It sets an uninfluential value on the regression for the formerly unavailable values while it makes the formerly incomplete lines now available for the estimations (see Table 4).

After the cleaning, the available sample includes no "NA's", and excludes the far outliers, keeping the core of each variable distribution unchanged and making all lines available for the estimations.

2.3. Model estimation

As per the standard Machine Learning, the cleaned dataset was randomly split, for each business sector, into two different files (see

Table 4
Variables distribution after replacing NA's with the mean value.

	Added value on production value	Depreciation and devaluation on costs	Payables to banks on current assets	Fixed asset coverage
Min	1.4	0.02	0.6	1.67
1st Quartile	24.3	0.8	35.9	56.5
Median	46.5	2.5	65.1	236.7
Mean	46.5	5.4	182.4	4316.1
3rd Quartile	66.7	5.4	114.1	2982.7
Max	100.0	65.6	6,965.9	146,302.4
NA's	0	0	0	0

Note: After replacing the NA's with the mean value, all observations are available for the estimations.

Table 5), namely a training set of around 75 % of the data and a testing set of the residual 25 %.

The training set was used for each business sector to tune the two models: a stepwise logistic regression and an XGBoost random forest estimation. Then, based on the resulting model, the testing set was used to verify the actual capability of each model to forecast defaults and to support the banking business.

It is worth highlighting that the differences between the training set and the testing set a way for mimicking an actual use, as the testing in this exercise would refer to new data coming from new observations, so generically different from the training set used for tuning the model, as it happens in the real use, in which the model is tuned by old data, and used for forecasting new cases.

2.4. *Logistic stepwise regression*

The logistic regression is performed on the available variables after the cleaning phase and on the cases included in the training set. It includes a set of ratios reporting different aspects of the firm activity and structure, including capitalization, debt structure, assets' structure, margins and returns, and some ratios about credit availability and use.

We performed a stepwise regression to select and concentrate the explanatory power into a smaller number of variables. The stepwise process selects the most significant variable, keeping the ones whose role does not significantly reduce the explanatory power of the variable set. This process must be conducted recursively, testing just one variable at a time, as the cross effects on the other variables of its exclusion cannot be simply estimated.

The results of the stepwise regression, reported in the Appendix, show that the stepwise process concentrates on a smaller list of variables, almost all statistically significant, and reports an informative value close to the complete list. Results also show that, as reported in the previous literature (Rikkers and Thibeault, 2011), different sectors report different coefficients and significance of variables due to the different balance sheet equilibria coming from their different business activity.

About the role of banking and balance sheet variables, from the tables in the Appendix we can see that banking variables resulted to be significant in 90 cases and reached the maximum significance in 15 cases. In comparison, balance sheet variables were significant in 128 cases, and reached the maximum significance in 9 cases. These results can be synthesized considering that banking variables concentrate significant information in a few variables, while balance sheet variables can add an essential contribution by including more variables, each one contributing with a smaller weight.

2.5. *XGBoost*

The following estimation is performed through the ensemble method based on a random forest called XGBoost.

As reported in the previous paragraphs, the random forest process is based on a set of trees in which each tree just sees a reduced random sample of the training set and a reduced random selection of the input variables. Thus, the first step of this process is in the hyperparameters tuning, determining the number of observations and variables for each tree, the number of trees, etc., to choose the optimal value for each.

The hyperparameters used by XGBoost are the maximum number of iterations for a tree generation (nrounds), the tree maximum depth (max_depht), the number of subsamples for each tree generation (subsample), the learning rate (eta), and the number of variables considered for each tree generation (colsample_bytree).

Following Gunnarsson et al. (2021), the possible suggested values for the hyperparameter are the ones reported in the following Table 6, along with the ones chosen by the tuning process for each business sector:

Then, for each sector, the random forest tuning was performed on the training set. No tables are available, as this kind of process is too complex to be described, so in some way, it is a black box, and its features can only be evaluated by the quality of its outcome performed on the testing set. This is one of the negative sides of many machine learning and artificial intelligence methods, which can reach higher performances using complex methods but are not understandable in terms of what drives their decision process. The actual, limited possibilities of evaluating the role of each input variable rely on a recursive process of verifying the differences in the outcome when one variable is excluded from the input list or by the effects the inclusion of a variable determines in the outcome, synthesized by some indexes, each measuring some aspects of the process and its outcome (the complete tables are reported in Appendix).

The "gain" score, measures the improvement in accuracy brought by each feature. In Table 7, for summarizing the results and verifying the roles of the two information sources we considered, namely banking variables and balance sheet variables, we just summed up the values reported by the banking variables and the balance sheet variables (first two rows). Then, (last two rows), we reported the values obtained by the first 10 variables in the ranking after summing it by source.

Results show that the total gain is mainly driven by the balance sheet variables, which report values ranging from 49.5 % to 63.9 %.

Table 5
Sample splitting.

Activity sectors	25	28	41	43	45	46	47
Training set	3,787	1,764	4,544	3,265	1,851	7,905	3,492
Testing set	1,232	605	1,536	1,149	622	2,636	1,147
Total	5,019	2,369	6,080	4,414	2,473	10,541	4,639

Note: For each sector, the testing set includes around 25 % of the total, randomly chosen.

Table 6
XGBoost hyperparameters tuning by activity sector.

Hyperparameters	Possible values	25	28	41	43	45	46	47
nrounds:	50; 100; 150	150	100	150	100	100	50	50
max.depth	1; 2; 3; 4	1	4	2	1	1	3	4
eta	0.2; 0.3; 0.4	0.4	0.4	0.4	0.3	0.3	0.4	0.2
subsample	0.6; 0.8	0.6	0.8	0.6	0.8	0.8	0.6	0.6
colsample_bytree	0.5; 0.75; 1	0.75	0.75	1	1	0.5	0.75	1

Table 7
XGBoost “gain” score by source and activity sector.

Gain	25	28	41	43	45	46	47
All banking variables	36.1 %	44.2 %	50.5 %	44.0 %	42.6 %	43.4 %	37.7 %
All balance sheet variables	63.9 %	55.8 %	49.5 %	56.0 %	57.4 %	56.6 %	62.3 %
Banking variables among the first 10	18.8 %	32.7 %	33.8 %	26.6 %	22.3 %	32.4 %	19.6 %
Balance sheet variables among the first 10	11.5 %	13.2 %	7.2 %	8.3 %	10.0 %	4.9 %	8.5 %

Note: For all sectors, the total gain is higher for the balance sheet variables, but when considering the top 10 variables, the higher share is for banking variables.

However, when just considering the first 10 variables, the banking ones score the highest values, doubling the balance sheet values.

The "cover" score (see Table 8) refers to the second-order derivative of the loss function with respect to the considered variable. Thus, it measures the variable potential impact on the loss function.

The “frequency” score (see Table 9) counts the number of times a feature is used in all generated trees, thus proxying the variable importance by the share of times it is included among the significant variables.

Both the "cover" and “frequency” scores confirm the same roles described above for the “gain” score, and previously obtained for the stepwise logistic, in which the balance sheet variables score a higher importance when considering the whole list, while the top variables’ analysis almost always report a clearly higher score for the banking variables.

3. Results

3.1. Estimation results

The following tables report the results of the stepwise logistic regression and XGBoost in terms of their capability to correctly forecast the defaults by comparing the actual defaults/no defaults to the fitted ones. For each model and each business sector, three settings are considered based on different cutoff values. The standard cutoff setting in discriminant analyses is 0.5, so the cases reporting a score higher than 0.5 are considered to default, while the values lower than this threshold are considered not to default. This is a standard setting for the general use of the statistical model, but the two error types have significantly different effects on banks.

Suppose a bank lends its money to a "good" customer. In that case, the expected gain is of the interest charged to it, typically lower than 10 %, so the lower income due to not financing a good customer is of this dimension order, while the loss in case the borrower defaults is around half the lending amount. Thus, the best balancing of the possible errors must be oriented to minimizing the default losses instead of equal balancing. Thus, we tested whether a 0.3 or a 0.2 cutoff could improve the bank’s expected income.

Based on these three cutoff values, the following tables report the results for each model as its confusion matrix (Table 10), i.e., the matrix reporting the correctly fitted actual defaults (true positive, TP), the incorrectly fitted as non-defaulters (false negative, FN), the incorrectly fitted as defaulters (false positive, FP) and the correctly fitted non-defaults (true negative, TN).

In the case of lending activity, no matter which human or computer-based model is considered, the most important values in the confusion matrix are the ones in the second line, i.e., the no-default fitted ones, as any bank would just lend to those borrowers. The FP is the bank’s nightmare, as each loan to an expectedly good borrower who turns out to be a defaulter determines a significant loss for the bank, whose value is set by the Basel II regulation to 45 % of the loan amount when no more precise estimation is available.

For each model and setting, after the table reporting the values, is also attached a second table reporting the statistical indexes of:

$Sensitivity = \frac{TP}{TP+FP}$(Correctly predicted defaults on total actual defaults);

$Specificity = \frac{TN}{TN+FN}$ (Correctly predicted no defaults on total actual no defaults);

$Positive\ predictive\ value = \frac{TP}{TP+FN}$ (Correctly predicted no defaults on total predicted no defaults);

$Negative\ predictive\ value = \frac{TN}{FP+TN}$ (Correctly predicted defaults on total predicted defaults);

$Correctly\ classified = \frac{TP+TN}{TP+FP+TN+FN}$ (Ratio of correctly classified cases on total cases).

While, in general, the correctly classified ratio is the most crucial index for assessing the model quality, for the lending activity, as the losses suffered in case of lending to a defaulter borrower are the key point, the capability to correctly forecast defaults is by far more important than the total classification correctness. Thus, the most significant parameter is sensitivity.

Table 8
XGBoost “cover” score by source and activity sector.

Cover	25	28	41	43	45	46	47
All banking variables	44.3 %	52.9 %	43.8 %	45.8 %	44.7 %	49.6 %	41.9 %
All balance sheet variables	55.7 %	47.1 %	56.2 %	54.2 %	55.3 %	50.4 %	58.1 %
Banking variables among the first 10	22.6 %	37.3 %	25.8 %	24.8 %	25.2 %	39.8 %	22.0 %
Balance sheet variables among the first 10	8.9 %	10.4 %	9.5 %	9.8 %	7.3 %	4.5 %	8.6 %

Note: For almost all sectors, just excluding the 28, the total cover is higher for the balance sheet variables. When considering the top 10 variables, the higher share is always and by far for banking variables.

Table 9
XGBoost “frequency” score by source and activity sector.

Frequency	25	28	41	43	45	46	47
All banking variables	28.5 %	29.7 %	36.2 %	30.2 %	34.1 %	31.4 %	29.7 %
All balance sheet variables	71.5 %	70.3 %	63.8 %	69.8 %	65.9 %	68.6 %	70.3 %
Banking variables among the first 10	9.7 %	19.0 %	16.4 %	13.3 %	15.1 %	18.6 %	14.9 %
Balance sheet variables among the first 10	9.8 %	11.7 %	8.4 %	8.6 %	9.4 %	5.3 %	7.4 %

Note: As for the previous tables, for all sectors, the total frequency is higher for the balance sheet variables, but when considering the top 10 variables, the higher share is almost always (excluding the 25) for banking variables.

Tables 11 and 12 report the stepwise logistic regression and XGBoost forecast indexes. While the correct classification rate reports seemingly interesting high values, ranging from 87.70 % to 95,54 % for the stepwise logistic and from 87.57 % to 96.86 % for XGBoost, these values are not satisfactory for direct use in a banking lending process.

Looking at the actual values (Tables 13–19), we can see that, e.g. for sector 41 (Table 15) even if 1306 out of 1459 no default cases are correctly forecast by the stepwise logistic model, scoring an excellent specificity of 97.32 % (1283 out of 1415, specificity of 95.60 % for the XGBoost), just 41 defaults out of 194 are correctly forecast, just reporting a sensitivity of 21.13 % (62 out of 194, sensitivity of 31.96 % for the XGBoost). As previously reported, the key point for a bank is the correct forecast of defaults, as lending to a defaulting firm will induce significant losses. Not lending to a good borrower just reduces the bank income of the (substantially lower) value of the charged interest.

Similar results are obtained for the other business sectors, reporting a typical high capability to forecast the good customers correctly, the negative predictive value ranging from 89.51 % to 97.47 % for the stepwise logistic, and from 90.67 % to 97.82 % for the XGBoost, while the capability to correctly forecast defaults (positive predictive value) ranges from 7.69 % to 53.85 % for the stepwise logistic and from 0 % to 68.75 % for the XGBoost.

The results reported in the appendix show that a different cutoff setting (to 0.2) can do a better balancing on the capability to correctly forecast defaults and no defaults, allowing, e.g., in sector 41 (Table A12), for a correct forecast of 131 out of 194 actual defaults, synthesized by a sensitivity score of 67.53 % (Table A17), at the cost of a specificity reduction to 82.19 % (115 out of 194, sensitivity of 59.28 %, specificity of 85.62 % for the XGBoost, see Table A18).

The Receiver Operating Characteristic (ROC) curves in Fig. 1 plot the differential characteristics of the two models’ outcomes in a graphical way, reporting the false positive rate (FPR) and true positive rate (TPR) for different cutoff settings. The closer the curve to the top left corner, the better the model classification performance.

Table 10
Confusion matrix.

		Actual		Total
		default	no default	
Fitted	default	TP	FN	TP+FN
	no default	FP	TN	FP+TN
	Total	TP+FP	FN+TN	TP+FN+FP+TN

Table 11
Stepwise logistic regression forecast indexes, cutoff 0.5.

	25	28	41	43	45	46	47
Sensitivity	5.56 %	6.25 %	21.13 %	9.52 %	10.00 %	5.43 %	10.29 %
Specificity	99.66 %	97.96 %	97.32 %	99.15 %	98.01 %	99.68 %	99.44 %
Positive predictive value	42.86 %	7.69 %	53.25 %	47.06 %	14.29 %	46.67 %	53.85 %
Negative predictive value	95.84 %	97.47 %	89.51 %	93.29 %	97.04 %	95.35 %	94.62 %
Correctly classified	95.54 %	95.54 %	87.70 %	92.60 %	95.18 %	95.07 %	94.16 %

Note: For all sectors, the sensitivity score is below 25 %, while the specificity score is above 95 %

Table 12
XGBoost forecast indexes, cutoff 0.5.

	25	28	41	43	45	46	47
Sensitivity	9.26 %	18.75 %	31.96 %	13.10 %	0.00 %	13.18 %	5.88 %
Specificity	99.15 %	98.98 %	95.60 %	99.53 %	98.67 %	99.08 %	98.70 %
Positive predictive value	33.33 %	33.33 %	51.24 %	68.75 %	0.00 %	42.50 %	22.22 %
Negative predictive value	95.97 %	97.82 %	90.67 %	93.56 %	96.74 %	95.69 %	94.33 %
Correctly classified	95.21 %	96.86 %	87.57 %	93.21 %	95.50 %	94.88 %	93.20 %

Note: The sensitivity score ranges from 0 % to 32 % %, while the specificity score is always above 95 %

Table 13
Stepwise logistic regression and XgBoost confusion matrices for sector 25, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	3	4	7	Fitted	default	5	10	15
	no default	51	1174	1225		no default	49	1168	1217
	Total	54	1178	1232		Total	54	1178	1232

Table 14
Stepwise logistic regression and XgBoost confusion matrices for sector 28, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	1	12	13	Fitted	default	3	6	9
	no default	15	577	592		no default	13	583	596
	Total	16	589	605		Total	16	589	605

Table 15
Stepwise logistic regression and XgBoost confusion matrices for sector 41, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	41	36	77	Fitted	default	62	59	121
	no default	153	1306	1459		no default	132	1283	1415
	Total	194	1342	1536		Total	194	1342	1536

Table 16
Stepwise logistic regression and XgBoost confusion matrices for sector 43, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	8	9	17	Fitted	default	11	5	16
	no default	76	1056	1132		no default	73	1060	1133
	Total	84	1065	1149		Total	84	1065	1149

Table 17
Stepwise logistic regression and XgBoost confusion matrices for sector 45, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	2	12	14	Fitted	default	0	8	8
	no default	18	590	608		no default	20	594	614
	Total	20	602	622		Total	20	602	622

Table 18
Stepwise logistic regression and XgBoost confusion matrices for sector 46, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	7	8	15	Fitted	default	17	23	40
	no default	122	2499	2621		no default	112	2484	2596
	Total	129	2507	2636		Total	129	2507	2636

Table 19
Stepwise logistic regression and XgBoost confusion matrices for sector 47, cutoff 0.5.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	7	6	13	Fitted	default	4	14	18
	no default	61	1073	1134		no default	64	1065	1129
	Total	68	1079	1147		Total	68	1079	1147

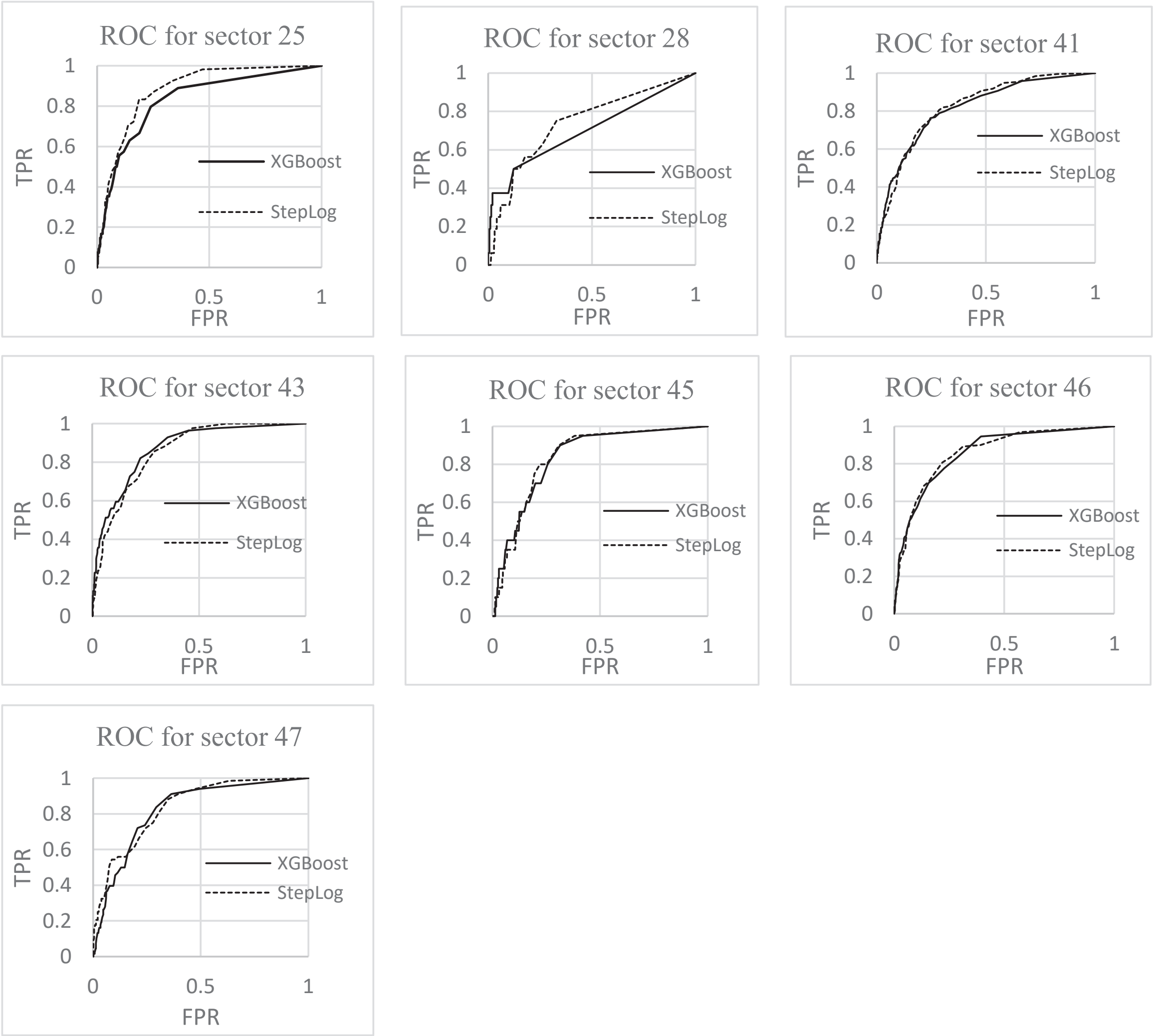

Fig. 1. ROC for XGBoost and Stepwise logistic, by activity sector.

Results show that just for sector 28, and partially for sector 25, XGBoost significantly overcomes stepwise logistic regression, while for the other sectors, the two curves are very close and often overlap each other.

The area under the ROC curve (AUC, see Table 20) is another way of measuring the models' features; the higher the AUC, the better the model, with a baseline reference value of 0.5, for which the results are equivalent to a random classification.

Also, the numeric quantification confirms the closeness of the two models, except for sector 28.

The progressive improvement obtained by a more rigorous selection of the possible borrower firms suggests that the same attitude can significantly improve the economic results for a bank, as the lower amount of default losses is more important than the income reduction due to not lending to the doubtful borrowers.

The main problem for optimizing the cutoff point is in correctly balancing the effect of losses due to unexpected defaults and gains from good borrowers.

3.2. Estimated income and efficiency

To verify the usefulness of the tested methods for banking and for possible optimization of results, we compared the models' outcomes in two more ways.

Firstly, we estimated the different models and settings of economic efficiency. To do this, we computed the expected income of the lending activity based on each model outcome. Considering that no bank would lend any money to any expected defaulting firm, the focus is on the actual results of the expected no-defaulting cases.

The expected profit can thus be computed as follows:

1) Positive case: The firm correctly refunded the loan. The bank income can be simply proxied as the unitary lending spread, i
2) Negative case: the firm defaults. The bank loss can be proxied as the unitary loss given default, LGD.

In formal terms, for each unitary loan j, the expected income $EI(j)$ of loan j can be computed as:

$$EI(j) = i_j \times (1 - PD_j)$$

Where:

PD_j is the probability to default of loan j

i_j is the unitary lending spread of loan j

While the corresponding expected loss $EL(j)$ can be computed as:

$$EL(j) = PD_j \times LGD_j$$

The expected net profit $E(j)$ can be thus computed as:

$$E(j) = EI(j) - EL(j) = i_j \times (1 - PD_j) - LGD_j \times PD_j$$

The expected income of loan j can then be compared among the different models and settings at any interest rate. So, referring to the confusion matrix in Table 10, and considering that no bank would lend to expected defaulters (TP+FN), the probability to default of the bank's loans portfolio PD can be computed[2] as the rate of actual defaults on the estimated good borrowers,

$$PD = \frac{FP}{FP + TN}$$

So, setting the interest rate spread to 5 %,[3] and LGD to 45 %, the standard level defined by the Basel II regulation when no more precise estimation is available, the bank's expected income EI can be computed as:

$$EI = 5\% \times (1 - PD) - 45\% \times PD$$

For a more complete and readable evaluation, we also considered two more possible extreme models: a perfect one, in which all firms were correctly classified (i.e. $PD = 0$), and a worst one, in which all cases were wrongly classified ($PD = 1$). On that base, we also estimated an efficiency level, which measures the relative position of the considered estimator in the linear scale from 0 % of the always wrong model to 100 % of the perfect forecasting model. Tables 21 to 27 report the results of these computations for the considered business sectors.

Tables 21 to 27 also report a further column named "Balancing spread," which reports the interest level needed for balancing losses and income (i.e., I such that $EI = 0$), evidencing the effect on lending spreads of the model quality, and the "Acceptance rate", which accounts for the share of loan requests positively classified by the model.

Results for the business sector 25, production of metal products, reported in Table 21, shows that both stepwise logistic regression

[2] In the simple case of unitary or equally sized loans.

[3] The average interest rate paid by the Italian building sector, as reported by Banca d'Italia was 4.92 % for investment-related operations and 5.3 % for liquidity needs, data referring to the end of 2022. Source: Banca d'Italia, Statistiche, Banche e istituzioni finanziarie: condizioni e rischiosità del credito, Table TRI30951 and TRI31100 from website www.bancaditalia.it.

Table 20
Stepwise logistic and XGBoost AUC by sector.

	25	28	41	43	45	46	47
Stepwise logistic	89.42 %	83.16 %	83.13 %	86.73 %	86.15 %	87.57 %	85.09 %
XGBoost	87.73 %	92.34 %	82.97 %	88.47 %	85.96 %	88.59 %	84.97 %

Note: The Area Under the Curve, AUC, shows that just for sector 28 XGBoost significantly overcomes stepwise logistic regression.

Table 21
Income and efficiency of different models and cutoff settings, sector 25.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0281	70.8 %	2.06 %	100.0 %
Stepwise Logistic	0.5	0.0290	72.2 %	1.95 %	99.4 %
Stepwise Logistic	0.3	0.0306	74.5 %	1.75 %	97.6 %
Stepwise Logistic	0.2	0.0315	75.8 %	1.58 %	95.5 %
XGBoost	0.5	0.0295	72.9 %	1.89 %	98.8 %
XGBoost	0.3	0.0304	74.2 %	1.76 %	97.2 %
XGBoost	0.2	0.0316	76.0 %	1.58 %	95.6 %
Best possible case		0.0478	100.0 %	0.00 %	95.6 %
Worst case		-0.0197	0.0 %		4.4 %

Note: Stepwise logistic and XGBoost report similar efficiency scores.

Table 22
Income and efficiency of different models and cutoff settings, sector 28.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0368	80.4 %	1.22 %	100.0 %
Stepwise Logistic	0.5	0.0365	79.9 %	1.17 %	97.9 %
Stepwise Logistic	0.3	0.0372	81.0 %	1.03 %	95.9 %
Stepwise Logistic	0.2	0.0373	81.2 %	0.97 %	94.4 %
XGBoost	0.5	0.0385	83.2 %	1.00 %	98.5 %
XGBoost	0.3	0.0396	85.0 %	0.86 %	97.4 %
XGBoost	0.2	0.0399	85.5 %	0.79 %	96.4 %
Best possible case		0.0487	100.0 %	0.00 %	97.4 %
Worst case		-0.0119	0.0 %		2.6 %

Note: For this sector, the XGBoost report efficiency scores around 4 % higher than stepwise logistic regression.

Table 23
Income and efficiency of different models and cutoff settings, sector 41.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		-0.0132	43.5 %	6.51 %	100.0 %
Stepwise Logistic	0.5	-0.0023	54.2 %	5.27 %	95.0 %
Stepwise Logistic	0.3	0.0099	66.4 %	3.77 %	84.8 %
Stepwise Logistic	0.2	0.0174	73.9 %	2.57 %	75.9 %
XGBoost	0.5	0.0031	59.6 %	4.63 %	92.1 %
XGBoost	0.3	0.0094	65.9 %	3.81 %	85.8 %
XGBoost	0.2	0.0143	70.7 %	3.09 %	79.9 %
Best possible case		0.0437	100.0 %	0.00 %	87.4 %
Worst case		-0.0568	0.0 %		12.6 %

Note: for this sector, results are different, reporting a better result for XGBoost for a cutoff set at 0.5, while stepwise logistic report higher efficiency when the cutoff is set at 0.2.

and XGBoost report an excellent efficiency, always higher than 70 %, slightly higher for the XGBoost, and subsequently nice balancing spread, ranging from 1,95 % of the stepwise logistic when the cutoff point is set at 0.5 (1.89 % for XGBoost), to a 1.58 % for both models when the cutoff is set at 0.2.

The same cutoff change reduces the acceptance rate (number of cases forecasted as not defaulters on total cases) from 99.4 % to 95.5 % for the stepwise logistic and from 98.8 % to 95.6 % for the XGBoost estimation.

Results for the business sector 28, production of machinery and equipment, reported in Table 22, shows a very low default rate, thus reporting higher efficiency rates (80.4 % for the whole sample), ranging from 79.9 % to 81.2 % for the logistic stepwise regression and even higher for the XGBoost model, ranging from 83.2 % to 85.5 %. The corresponding balancing spreads are very low, 1.17–0.97 % for the logistic and 1.00–0.79 % for the XGBoost. The XGBoost model reports better performances in all indicators for this business sector, showing higher income and efficiency, lower balancing spreads, and higher acceptance rates.

Table 24
Income and efficiency of different models and cutoff settings, sector 43.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0134	58.5 %	3.53 %	100.0 %
Stepwise Logistic	0.5	0.0162	61.9 %	3.24 %	98.5 %
Stepwise Logistic	0.3	0.0199	66.6 %	2.79 %	95.5 %
Stepwise Logistic	0.2	0.0246	72.5 %	2.14 %	90.1 %
XGBoost	0.5	0.0175	63.6 %	3.10 %	98.6 %
XGBoost	0.3	0.0230	70.5 %	2.46 %	95.6 %
XGBoost	0.2	0.0262	74.5 %	2.04 %	92.3 %
Best possible case		0.0463	100.0 %	0.00 %	92.7 %
Worst case		-0.0329	0.0 %		7.3 %

Note: For this sector, the XGBoost report efficiency scores slightly higher than stepwise logistic regression.

Table 25
Income and efficiency of different models and cutoff settings, sector 45.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0339	77.0 %	1.50 %	100.0 %
Stepwise Logistic	0.5	0.0344	77.7 %	1.37 %	97.7 %
Stepwise Logistic	0.3	0.0339	77.0 %	1.33 %	95.2 %
Stepwise Logistic	0.2	0.0352	79.0 %	1.12 %	92.9 %
XGBoost	0.5	0.0333	76.0 %	1.52 %	98.7 %
XGBoost	0.3	0.0361	80.4 %	1.16 %	96.3 %
XGBoost	0.2	0.0364	80.9 %	0.96 %	92.1 %
Best possible case		0.0484	100.0 %	0.00 %	96.8 %
Worst case		-0.0145	0.0 %		3.2 %

Note: For this sector, both models report efficiency scores just slightly higher than the whole sample.

Table 26
Income and efficiency of different models and cutoff settings, sector 46.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0255	68.3 %	2.29 %	100.0 %
Stepwise Logistic	0.5	0.0266	69.8 %	2.20 %	99.4 %
Stepwise Logistic	0.3	0.0286	72.8 %	1.94 %	97.5 %
Stepwise Logistic	0.2	0.0306	75.6 %	1.66 %	94.9 %
XGBoost	0.5	0.0280	71.9 %	2.03 %	98.5 %
XGBoost	0.3	0.0307	75.8 %	1.70 %	96.7 %
XGBoost	0.2	0.0318	77.4 %	1.50 %	94.0 %
Best possible case		0.0476	100.0 %	0.00 %	95.1 %
Worst case		-0.0220	0.0 %		4.9 %

Note: For this sector, the XGBoost report efficiency scores slightly higher than stepwise logistic regression.

Table 27
Income and efficiency of different models and cutoff settings, sector 47.

	Cutoff	Income	Efficiency	Balancing spread	Acceptance rate
Whole sample		0.0204	63.8 %	2.84 %	100.0 %
Stepwise Logistic	0.5	0.0228	67.2 %	2.56 %	98.9 %
Stepwise Logistic	0.3	0.0258	71.2 %	2.18 %	96.1 %
Stepwise Logistic	0.2	0.0277	73.8 %	1.86 %	92.1 %
XGBoost	0.5	0.0213	65.1 %	2.70 %	98.4 %
XGBoost	0.3	0.0237	68.4 %	2.38 %	95.4 %
XGBoost	0.2	0.0262	71.7 %	2.04 %	92.4 %
Best possible case		0.0470	100.0 %	0.00 %	94.1 %
Worst case		-0.0267	0.0 %		5.9 %

Note: For this sector, the stepwise logistic regression report efficiency scores slightly higher than XGBoost.

Results for the business sector 41, Construction of buildings, reported in Table 23, show higher default rates (whole sample efficiency of 43.5 %) but an excellent capability of models to explain the following defaults. In this case, the efficiency ranges from 54.2 % for the logistic, with a cutoff set at 0.5–73.9 % for the stricter 0.2 cutoff (59.6–70.7 % for the XGBoost). The efficiency significant gain obtained by the cutoff stricter setting is reflected in the balancing spread reduction, halved from 5.27 % to 2.57 % for the stepwise logistic and from 4.63 % to 3.09 % for the XGBoost.

These differences are reflected by significant reductions in the acceptance rate, of more than 12 % for XGBoost and around 20 % for the stepwise logistic model.

Interestingly, in this case, the 0.5 cutoff reports nicer results for the XGBoost, while the 0.2 setting reverses the ranking, showing better efficiency and income and lower balancing spreads for the logistic.

Also, the results for the business sector 43, Specialized construction works, reported in Table 24, show an initially limited capability to capture the default signals, reporting an efficiency just slightly higher than the whole sample (58.5 %) for the 0.5 cutoff setting on both models (61.9 % for the logistic stepwise regression, and 63.6 % for XGBoost), but signaling some better result for XGBoost. A significant role is for the cutoff setting, as in both models we obtain a clearly higher performance setting the cutoff at 0.2, showing efficiency values of 72.5 % for the logistic stepwise regression, and of 74.5 % for the XGBoost model.

The corresponding balancing spreads lowers from 3.24 % to 2.14 % for the logistic model and from 3.10 % to 2.04 % for the XGBoost. Even in the stricter cutoff case, the acceptance rates are higher than 90 % for both models.

For sector 45, wholesale and retail of cars and motorcycles, reported in Table 25, results show an initial low capability to efficiently select the possible borrowers, so that the whole sample reference efficiency of 77 % is just slightly improved by the stepwise logistic model (77.7 %), and is even worse for XGBoost, whose efficiency results of 76 % for the 0.5 cutoff setting. Here, the stricter selection obtained when setting the cutoff at 0.2, which lowers the acceptance rate from 97.7 % and 98.7 % of the logistic and XGBoost models, respectively, to 92.9 % and 92.1 %, is just capable of raising the efficiency of some points, but reversing the ranking and showing the top result for XGBoost (80.9 %), 1.9 % higher than the logistic regression.

Results for the business sector 46, Wholesale (excluding cars and motorcycles), reported in Table 26, show some better initial results for the 0.5 cutoff setting, slightly improving the efficiency of the whole sample, which increases, respectively of 75.6 % and 77.4 % for the logistic and XGBoost with 0.2 cutoff setting, coupled with a limited selection of around 94 % for both.

In this sector, the XGBoost reports slightly better results on all indexes and cutoff settings.

Finally, the results for the business sector 47, retail (excluding cars and motorcycles), reported in Table 27, show some exciting capability to capture the default risk, signaled by the higher efficiency of both models in comparison with the whole sample (63.8 %) already from the 0.5 cutoff setting (67.2 % for logistic and 65.1 % for XGBoost), and gaining more efficiency when setting the cutoff at the stricter level of 0.2 (73.8 %for logistic and 71.7 % for XGBoost).

In this sector, is the logistic model to score higher performances on all indicators and for all cutoff settings.

4. Discussion and conclusions

The Llewellyn (2012) analysis, reporting that "*Bank business models are not static and evolve over time and under the influence of a complex mix of exogenous and endogenous pressures,"* sounds particularly capable of describing the banks' business models' evolution in recent years. Among the evolution drivers, machine learning and AI recently gained the stage for the possible technological jumps they can bring in the lending decision process evolution, even if the attention devoted to confirming the effectiveness and efficiency of these methods in real banking use, has not been yet sufficient.

To add more information on this topic, in this paper, we tested two different discriminant models based on a stepwise logistic regression and the XGBoost random forest model on a set of 28 banking variables and 55 balance sheet ratios. The models are tested on a sample of 35,535 cases from 7 different business sectors, of which, for each sector, around 75 % were used for training the models, and the residual 25 % for testing its capability to forecast defaults and non-defaults correctly.

With this aim, we developed an efficiency index for measuring each model's capability to correctly select the good borrowers, balancing the different effects of refusing the loan to a good customer and lending to a defaulter. Also, we computed the balancing spread to quantify the different models' efficiency in terms of credit costs for the lenders.

The results of the different models and estimates show that, even if different sectors report different results, the two models have similar forecasting capabilities and that the cutoff setting, as to say, the selection level chosen in the actual use of the scoring models, can be the decisive point for the effective use of these techniques. The sensitivity levels, estimated income, and balancing spreads show that the banking efficiency is more affected by this parameter setting than the model choice. Results show a slight improvement when considering the standard cutoff setting of 0.5, which is not sufficient for an economically favorable direct use. However, the selection capabilities interestingly improve when a more rigorous selection of the possible borrowers limits the actual lending to the best ones, setting the cutoff at the 0.2 level, and this is without a significant reduction in lending volumes, so in the substantial role of financing the real economy.

These results allow for an upbeat assessment of the tested quantitative models, as the information included in these variables allows for a nice selection of the good borrowers, at least for the considered SMEs category, for which an automated process, characterized by low costs, decision speed and repeatability, can adequately match with the high number of requests, limited amounts, and high diversification.

Implicitly, the results also remind us that the credit scoring models' features are so good when performing the estimations separately by activity sector and for dimensionally coherent firms' categories, as repeatedly confirmed by the previous literature.

These results are actually valuable for the fintech case, where there is no direct contact among banks and borrowers, and can possibly improve by adding some more information sources, e.g. the previous lending results and macro indicators.

Nevertheless, the quantitative models to be used must be attentively selected and tuned to optimally choose the most significant variables for each activity sector, the cutoff settings, and to correctly balance the effects of rejecting the credit request to a good borrower and of lending to a defaulter.

With these attentions, the use of quantitative models can allow banks to adapt their business model to the technology improvements, can bring a reduction in personnel costs, faster answers, and evaluations' homogeneity, lower spreads, and a better credit allocation, which are fundamental for supporting the real economy sustainable growth.

Data availability

The data that has been used is confidential.

Acknowledgements

This research is funded by Fondazione di Sardegna.

Appendix A

Table A1
Variables list.

Balance sheet variables	Description
Other revenues on production value	Other revenues / production value
Intangible asset on production value	Intangible asset / production value
Immediate liquidity on total assets	Immediate liquidity / total assets
Added value on production value	Added value / production value
Depreciation and devaluation on costs	Depreciation and devaluation / total costs
Financial autonomy	Equity / total liabilities + equity
Payables to banks on current assets	Payables to banks / current assets
Cash flow on production value	Cash flow / production value
Unit cash flow (on total revenues)	Cash flow / total revenues
Net assets coverage	(Equity+ long term liabilities) / net assets
Fixed asset coverage	(Equity+ long term liabilities) / fixed assets
Financial coverage index	(Equity+ long term liabilities) / financial assets
Labor cost on revenues	Labor costs / revenues
Labor cost on production value	Labor cost / production value
Unit labor cost	Labor cost / sales
Credits on Total assets	Credits / Total assets
Current assets / current liabilities	Current assets / current liabilities
Short-term payables on amounts due to banks	Short-term payables / amounts due to banks
Short-term payables on Net worth	Short-term payables / Net worth
Payables on short-term debts	Payables / short-term debts
Debts on Net worth	Debts / Net worth
Payables to suppliers on Net worth (shareholders' equity)	Payables to suppliers / shareholders' equity
Payables to suppliers on Total debt	Payables to suppliers / Total debt
Inventory duration	(Inventory / sales) *360
Degree of indebtedness	Total debt / (total liabilities + equity)
Intangible assets on shareholders' equity	Intangible assets / shareholders' equity
Intangible assets on Total assets	Intangible assets / total assets
Tangible assets on shareholders' equity	Tangible assets / shareholders' equity
Tangible assets on Total assets	Tangible assets / total assets
Short-term debt	Short-term debt / total debt
Debt to banks	Debt to banks / total debt
Financial dependence index	Financial dependence index
Debt burden index	Financial costs / EBITDA
Index of rigidity of assets	Index of rigidity of assets
EBITDA on revenues	EBITDA / revenues
Financial interest on revenues	Financial interest costs / revenues
Shareholders' equity on (long-term equity and payables)	Shareholders' equity / (long-term equity and payables)
Shareholders' equity on equity and inventories	Shareholders' equity / equity and inventories
Leverage	Leverage
Gross operating profitability	EBITDA / production value
Inventories on Total assets	Inventories / total assets
Inventories on Short-term debt	Inventories / short-term debt
Inventories on Total debt	Inventories / total debt
Inventories on Bank debt	Inventories / bank debt
ROA	EBIT / total assets
ROD	Interest costs / total debt
ROE	Net worth / shareholders' equity
ROI	EBIT / invested capital
ROS	EBITDA / total sales
Invested capital turnover	Sales / invested capital

(*continued on next page*)

Table A1 (*continued*)

Balance sheet variables	Description
Working capital turnover	Sales / working capital
Inventory turnover	Sales / Inventory
Value of production on Total Assets	Production value / total assets
Added value on Revenues	Added value / Revenues
Production value on Inventories	Production value / Inventories
Banking relationship variables	**Description**
Annual used credit on granted credit	Annual used credit / granted credit
Annual total credit amounts on debit amounts	Annual total credit amounts / debit amounts
Annual overdraft on granted credit	Annual overdraft / granted credit
Annual overdraft on granted credit (compensated)	Annual overdraft / granted credit (compensated)
Annual average of debit movements	Annual amount of debit movements / number of debit movements
Annual total credit amounts on debit/credit amounts	Annual total credit amounts / total debit + credit amounts
Annual average of credit movements	Annual amount of credit movements / number of credit movements
Percentage of debit movements out of total number of annual debit/credit movements	Percentage of debit movements / total number of annual debit/credit movements
Highest value of unpaid checks at first presentation per year	Highest value of unpaid checks at first presentation per year
Highest value of unpaid checks per year	Highest value of unpaid checks per year
Average value of checks unpaid at first presentation per year	Amount of unpaid checks at first presentation per year / number of unpaid checks at first presentation
Average value of unpaid checks per year	Amount of unpaid checks per year / number of unpaid checks
Total of checks unpaid at first presentation per year	number of checks unpaid at first presentation per year
Total unpaid checks per year	Total number of unpaid checks per year
Violation months	Number of months of credit limit violation over the year
Violation months (compensated)	Number of months of credit limit violation (compensated) over the year
Credit limit violation flag	Credit limit violation over the year (flag)
Credit limit violation (compensated) flag	Compensated credit limit violation over the year (flag)
Semiannual Violation months	Number of months of credit limit violation in the last 6 months
Semiannual Violation months (compensated)	Number of months of credit limit violation (compensated) in the last 6 months
Semiannual Credit limit violation flag	Credit limit violation in the last 6 months (flag)
Semiannual Credit limit violation (compensated) flag	Compensated credit limit violation in the last 6 months (flag)
Annual used credit on granted credit	Average annual used credit / granted credit
Value of unpaid checks on revenues	Value of unpaid checks / revenues
Annual total credit/debit amounts on revenues	Annual total credit + debit amounts / revenues

Table A2
Stepwise regression results per business sector.

	25		28		41		43		45		46		47	
(Intercept)	-6.488e+00	***	-1.054e+01	***	-5.361e+00	***	-6.484e+00	***	-6.304e+00	***	-5.543e+00	***	-4.431e+00	***
Annual used credit on granted credit	8.963e-03	**			1.203e-02	***	8.901e-03	**	2.774e-02	***	9.172e-03	***	6.872e-03	.
Annual total credit amounts on debit amounts	-9.404e-03	.									1.103e-02	.		
Annual overdraft on granted credit					-1.462e-01	.								
Annual overdraft on granted credit (compensated)					1.399e-01	.			-7.686e-02	**				
Annual average of debit movements									-3.826e-07					
Annual total credit amounts on debit/ credit amounts					6.940e-03	*					-5.705e-02	*		
Annual average of credit movements									7.081e-07	.				
Percentage of debit movements out of total number of annual debit/credit movements							1.033e-02		2.082e-02	.	6.651e-03	.		
Highest value of checks unpaid at first presentation per year	8.680e-05	**	1.428e-04	*	3.435e-05	*			6.536e-05	*	1.119e-05	*	9.948e-06	*
Highest value of unpaid checks per year			3.021e-03						4.596e-04				6.964e-05	.

(*continued on next page*)

Table A2 (*continued*)

	25		28		41		43		45		46		47	
Average value of checks unpaid at first presentation per year	-5.269e-04	*	-8.609e-04		-1.849e-04	*			-2.581e-04	*				
Average value of unpaid checks per year									-3.752e-03				-5.128e-04	
Total of checks unpaid at first presentation per year	7.670e-02	.	1.358e-01											
Total unpaid checks per year			-1.502e+01		4.250e-01	**					2.880e-01	**	1.225e-01	.
Annual number of months of overdraft	2.034e-01	*	3.840e-01	.					6.295e-01	**	-1.919e-01			
Annual number of months of overdraft (compensated)	-2.159e-01	.	-6.248e-01	*					-6.215e-01	*	2.480e-01	.		
Annual overdraft presence flag (compensated)	5.310e-01	.			4.206e-01	*					1.767e+00	***	1.019e+00	***
Annual overdraft presence flag			1.383e+00	.			6.721e-01	*			-1.149e+00	*		
Number of months of overdraft every six months					4.768e-01	***					7.448e-01	***		
Number of months of overdraft (compensated) every six months	2.840e-01	*	8.448e-01	**			3.120e-01	*			-5.760e-01	*		
Overdraft presence flag every six months			2.076e+00	.					6.545e-01	.	-1.322e+00	*		
Annual continuous months of overdraft (compensated)			-1.816e+00								1.308e+00	*		
Annual continuous months of overdraft							-2.718e-01		-7.791e-01	*				
Continuous months of overdraft per year							4.928e-01	*	9.996e-01	**				
Continuous months of overdraft every six months			-9.048e-01	.	-1.780e-01	*	5.187e-01	.					2.547e-01	***
Continuous months of overdraft (compensated) every six months			8.356e-01				-7.628e-01	**						
Other revenues on production value			5.756e-02	*					-4.323e-02	.	1.499e-02			
Intangible asset on production value			2.930e-02	*			4.316e-02	**						
Immediate liquidity on total assets	-4.396e-02				-3.614e-02	**								
Added value on production value	-1.250e-02	*					-2.631e-02	**	-1.079e-01	.			-1.789e-02	
Depreciation and devaluation on costs							-7.133e-02	*						
Financial autonomy							-1.714e-01							
Payables to banks on current assets									-4.600e-03	*			2.520e-03	**
Cash flow on production value											-8.715e-02	.		
Unit cash flow (on total revenues)	-5.872e-02	*	-6.573e-02	.							8.023e-02	.		
Net active coverage					-1.865e-02	***	1.669e-01		3.262e-02	.			-1.924e-02	*
Fixed asset coverage	1.511e-03	*												
Financial coverage index	-5.054e-04	.												
Labor cost on revenues							3.078e-02	***	1.203e-01	.				
Labor cost on production value											2.132e-03		-2.613e-03	.
Unit labor cost									-1.535e-01	.			2.187e-02	
Credits on Total assets	1.720e-02	*							2.470e-02	*				

(*continued on next page*)

Table A2 (*continued*)

	25		28		41		43		45		46		47	
Current assets / current liabilities							-1.178e-03		-1.060e-02	.	1.121e-03	*		
Short-term payables on amounts due to banks			3.279e-04	*										
Short-term payables on Net worth			6.742e-04	**	-3.077e-05	.	-2.908e-04	**	4.286e-04	***				
Payables on short-term debts							7.490e-04	**	-4.438e-03	*				
Debts on Net worth														
Payables to suppliers on Net worth (shareholders' equity)			-7.697e-04	*	1.619e-04	*								
Payables to suppliers on Total debt					5.078e-03	.								
Inventory duration	5.190e-03	*					-8.656e-04	.			1.551e-03	***		
Degree of indebtedness	2.206e-02	*	4.858e-02	**			2.291e-02	**	3.881e-02	**	1.646e-02	***		
Intangible assets on shareholders' equity	1.225e-03	*	-1.687e-03											
Intangible assets on Total assets			-7.378e-02	.			-5.882e-02	*	5.656e-02	**				
Tangible assets on shareholders' equity	-4.590e-04	*					-3.151e-04		-6.449e-04	.	-2.769e-04	.	3.521e-04	**
Tangible assets on Total assets	1.578e-02	.	-3.924e-02	*	-1.187e-02	***	-1.070e-02						-1.394e-02	**
Short-term debt			-1.418e-02	.					-3.956e-02	**				
Debt to banks					7.264e-03	**								
Financial dependence index	6.310e-03	*	-7.777e-03	.			-2.294e-03	*						
Debt burden index									-9.095e-03	*				
Index of rigidity of assets			3.323e-02	.										
EBITDA on revenues			9.486e-02	**	6.476e-05	.			7.505e-02	*	-2.560e-02	.		
Financial inerest on revenues									1.607e-01	**				
Shareholders' equity on (long-term equity and payables)							-5.433e-03	.						
Shareholders' equity on equity and inventories	1.210e-02	*	2.566e-02	**	7.788e-03	***							9.624e-03	*
Leverage							2.988e-04	***						
Gross operating profitability (EBITDA on production value)			-6.930e-02	.			2.337e-02	*	-9.164e-02	*	4.341e-02	**		
Inventories on Total assets	6.896e-02	*	2.391e-02	.			-3.241e-02		3.659e-02					
Inventories on Short-term debt							-2.087e-03	*			-1.790e-03	*		
Inventories on Total debt	-4.944e-02	*					2.950e-02	.	-4.514e-02	**				
Inventories on Bank debt			-2.367e-03	*										
ROA					-1.677e-02	.					-2.255e-02	*	-2.704e-02	***
ROD			1.834e-01											
ROE	-2.753e-03	.	-4.495e-03	*							-2.294e-03	*		
ROI	-2.605e-02	*							-2.774e-02				-1.650e-02	.
ROS													2.975e-02	*
Capital invested rotation			2.540e-02										-3.116e-03	*
Working capital turnover									2.409e-03	*			-7.730e-04	.
Inventory turnover	-1.137e-04	.	2.063e-03		-2.203e-05				4.207e-03	.				
Value of production on Total Assets	-5.332e-03		-2.162e-02		-3.457e-03	**			-6.925e-03	*				
Added value on Revenues									9.406e-02	.				
Production value on Inventories			-2.304e-03	.					-4.666e-03	*				
Annual used credit on granted credit	6.550e-01	.	1.408e+00	*	1.509e+00	***	8.504e-01	**	1.462e+00	*	1.111e+00	***	1.278e+00	***
Value of unpaid checks on revenues							1.833e+03	***						
Annual total credit/debit amounts on revenues	-2.196e+00	*	-2.407e+00	.	-4.750e-05	*	-2.029e+00	***	-2.604e+00	*	-2.584e+00	***	-2.838e+00	***

Table A3
XGBoost importance scores for sector 25.

	Gain	Cover	Frequency
Annual used credit on granted credit	7.2 %	8.4 %	3.4 %
Annual total credit/debit amounts on revenues	4.1 %	3.4 %	3.5 %
Number of months of overdraft (compensated) every six months	2.9 %	4.2 %	0.5 %
Degree of indebtedness	2.5 %	2.2 %	1.6 %
Financial dependence index	2.4 %	1.3 %	2.4 %
Annual overdraft on granted credit	2.4 %	2.4 %	1.4 %
Added value on production value	2.4 %	2.1 %	1.7 %
Annual number of months of overdraft (compensated)	2.3 %	4.3 %	0.9 %
ROA	2.2 %	2.3 %	1.9 %
Immediate liquidity on total assets	2.0 %	1.0 %	2.1 %
Annual average of credit movements	2.0 %	1.2 %	2.0 %
Credits on Total assets	1.9 %	2.5 %	1.6 %
Annual total credit amounts on debit amounts	1.8 %	0.9 %	2.0 %
Acid test	1.7 %	1.7 %	1.8 %
Unit labor cost	1.7 %	0.8 %	1.8 %
Annual used credit on granted credit	1.6 %	1.2 %	2.4 %
ROE	1.6 %	1.8 %	1.9 %
Payables to suppliers on Total debt	1.6 %	1.5 %	2.5 %
Annual total credit movements on debit movements	1.6 %	1.4 %	1.8 %
Payables to suppliers on Net worth (shareholders’ equity)	1.5 %	0.6 %	1.6 %
Debt to banks	1.5 %	1.8 %	1.5 %
Production value on Inventory	1.5 %	0.9 %	1.1 %
ROS	1.5 %	2.3 %	1.7 %
Other revenues on production value	1.5 %	1.0 %	1.8 %
Working capital turnover	1.4 %	0.8 %	1.4 %
Tangible assets on Total assets	1.4 %	1.2 %	1.3 %
Payables to banks on current assets	1.4 %	1.2 %	1.5 %
Index of rigidity of assets	1.4 %	1.1 %	1.1 %
Financial coverage index	1.3 %	1.1 %	1.5 %
medium and long term debt indebtness	1.3 %	0.9 %	1.4 %
Annual average of credit/debit movements	1.3 %	0.9 %	1.4 %
Financial interest on Added value	1.3 %	3.6 %	1.6 %
EBITDA on revenues	1.3 %	0.6 %	1.3 %
Tangible assets on shareholders’ equity	1.2 %	0.6 %	1.1 %
Inventories on Short-term debt	1.2 %	1.1 %	1.1 %
Number of months of overdraft every six months	1.2 %	1.9 %	0.6 %
ROI	1.2 %	0.8 %	1.5 %
Annual average of debit movements	1.2 %	0.8 %	1.0 %
Inventory duration	1.2 %	0.7 %	1.1 %
Depreciation and devaluation on costs	1.1 %	0.5 %	1.6 %
Intangible assets on Total assets	1.1 %	0.7 %	1.5 %
Unit cash flow (on total revenues)	1.0 %	0.8 %	1.1 %
Annual number of months of overdraft	1.0 %	1.3 %	0.5 %
Current assets / current liabilities	0.9 %	0.5 %	1.4 %
Labor cost on production value	0.9 %	0.5 %	1.2 %
Continuous months of overdraft (compensated) every six months	0.9 %	1.6 %	0.7 %
Inventories on Total debt	0.9 %	0.5 %	1.2 %
Capital invested rotation	0.9 %	0.8 %	1.1 %
Annual total credit amounts on debit/credit amounts	0.9 %	0.6 %	1.2 %
Immediate liquidity	0.8 %	0.4 %	0.9 %
Inventories on Bank debt	0.8 %	0.9 %	1.0 %
Shareholders’ equity on equity and inventories	0.8 %	1.5 %	1.0 %
Fixed asset coverage	0.8 %	0.7 %	0.9 %
Added value on Revenues	0.7 %	0.7 %	0.7 %
Cash Flow on production value	0.7 %	0.9 %	1.2 %
Short-term debt	0.7 %	1.1 %	0.9 %
Average value of checks unpaid at first presentation per year	0.7 %	3.1 %	0.6 %
Annual overdraft on granted credit (compensated)	0.7 %	0.5 %	1.0 %
Total of checks unpaid at first presentation per year	0.7 %	2.4 %	0.5 %
Labour cost on revenues	0.7 %	0.4 %	0.9 %
ROD	0.7 %	0.6 %	1.0 %
Financial interest on revenues	0.6 %	0.4 %	1.0 %
Short-term payables on Net worth	0.6 %	0.5 %	0.9 %
Financial autonomy	0.6 %	0.2 %	0.5 %
Annual continuous months of overdraft	0.6 %	0.6 %	0.9 %
Inventories on Total assets	0.6 %	0.3 %	1.0 %
Value of production on Total Assets	0.6 %	0.7 %	0.8 %

(*continued on next page*)

Table A3 (*continued*)

	Gain	Cover	Frequency
Payables on short-term debts	0.6 %	0.4 %	0.9 %
Intangible assets on shareholders' equity	0.6 %	0.4 %	1.1 %
Highest value of checks unpaid at first presentation per year	0.5 %	2.9 %	0.9 %
Leverage	0.5 %	0.7 %	0.6 %
Net active coverage	0.5 %	0.4 %	0.5 %
Shareholders' equity on (long-term equity and payables)	0.4 %	0.3 %	0.5 %
Gross operating profitability (EBITDA on production value)	0.4 %	0.4 %	0.5 %
Debts on Net worth	0.4 %	0.2 %	0.7 %
Short-term payables on amounts due to banks	0.4 %	0.2 %	0.7 %
Intangible asset on production value	0.2 %	0.2 %	0.5 %
Inventory turnover	0.2 %	1.0 %	0.5 %
Continuous months of overdraft every six months	0.2 %	0.2 %	0.5 %
Average value of unpaid checks per year	0.2 %	0.1 %	0.1 %
Inventories on medium and long term debt	0.2 %	0.3 %	0.3 %
Percentage of debit movements out of total number of annual debit/credit movements	0.2 %	0.1 %	0.3 %
Highest value of unpaid checks per year	0.1 %	0.1 %	0.2 %

Table A4
XGBoost importance scores for sector 28.

	Gain	Cover	Frequency
Annual used credit on granted credit	8.8 %	8.1 %	4.1 %
Highest value of checks unpaid at first presentation per year	7.6 %	11.8 %	3.8 %
Annual total credit/debit amounts on revenues	4.7 %	3.0 %	3.4 %
Annual number of months of overdraft (compensated)	4.2 %	4.8 %	1.7 %
Shareholders' equity on equity and inventories	4.2 %	4.5 %	2.8 %
Annual overdraft on granted credit (compensated)	4.1 %	3.3 %	2.8 %
Credits on Total assets	3.6 %	2.5 %	4.5 %
Annual used credit on granted credit	3.3 %	6.4 %	3.1 %
ROA	2.8 %	2.2 %	2.8 %
Index of rigidity of assets	2.6 %	1.2 %	1.7 %
Intangible asset on production value	2.2 %	1.6 %	2.4 %
Annual average of credit/debit movements	2.2 %	1.7 %	2.1 %
Number of months of overdraft (compensated) every six months	2.1 %	2.2 %	1.0 %
Cash Flow on production value	2.0 %	1.7 %	1.7 %
Degree of indebtedness	2.0 %	2.8 %	2.1 %
Current assets / current liabilities	1.9 %	1.1 %	2.1 %
Fixed asset coverage	1.9 %	1.8 %	2.4 %
Unit cash flow (on total revenues)	1.7 %	1.5 %	2.4 %
Payables to banks on current assets	1.7 %	1.3 %	2.1 %
Financial dependence index	1.6 %	1.5 %	2.1 %
Added value on production value	1.6 %	1.1 %	2.4 %
medium and long term debt indebtness	1.5 %	1.3 %	1.7 %
Total unpaid checks per year	1.3 %	2.3 %	0.7 %
Annual overdraft on granted credit	1.3 %	3.1 %	1.4 %
Short-term payables on amounts due to banks	1.3 %	1.2 %	2.1 %
Added value on Revenues	1.2 %	0.8 %	1.4 %
Financial interest on revenues	1.2 %	0.9 %	1.0 %
Working capital turnover	1.2 %	0.8 %	1.7 %
Inventories on Total assets	1.1 %	0.6 %	1.4 %
Gross operating profitability (EBITDA on production value)	1.1 %	0.9 %	1.4 %
Other revenues on production value	1.1 %	0.9 %	1.7 %
Inventories on Bank debt	1.0 %	0.7 %	1.4 %
Annual total credit movements on debit movements	1.0 %	0.8 %	1.4 %
Immediate liquidity	1.0 %	0.7 %	1.4 %
ROI	1.0 %	0.9 %	1.4 %
Production value on Inventory	1.0 %	0.6 %	1.0 %
ROE	0.8 %	1.0 %	1.0 %
ROS	0.8 %	0.6 %	1.4 %
Tangible assets on Total assets	0.8 %	1.1 %	2.1 %
Immediate liquidity on total assets	0.8 %	0.7 %	1.4 %
Depreciation and devaluation on costs	0.7 %	0.5 %	1.0 %
Payables to suppliers on Total debt	0.7 %	0.3 %	0.7 %
Payables on short-term debts	0.7 %	0.8 %	1.4 %
Inventories on Total debt	0.7 %	0.9 %	1.4 %
Annual total credit amounts on debit amounts	0.6 %	0.3 %	0.7 %
Financial coverage index	0.6 %	0.3 %	0.7 %

(*continued on next page*)

Table A4 (*continued*)

	Gain	Cover	Frequency
Intangible assets on Total assets	0.6 %	0.7 %	1.0 %
Annual continuous months of overdraft	0.5 %	3.5 %	0.7 %
Financial interest on Added value	0.5 %	0.4 %	0.7 %
Percentage of debit movements out of total number of annual debit/credit movements	0.5 %	0.4 %	0.7 %
Inventories on Short-term debt	0.5 %	0.5 %	0.7 %
Annual number of months of overdraft	0.5 %	0.0 %	0.3 %
Tangible assets on shareholders' equity	0.5 %	0.3 %	0.7 %
Number of months of overdraft every six months	0.4 %	0.4 %	0.3 %
Labour cost on revenues	0.4 %	0.6 %	0.7 %
Debts on Net worth	0.4 %	0.6 %	0.7 %
Annual total credit amounts on debit/credit amounts	0.4 %	0.1 %	0.3 %
EBITDA on revenues	0.3 %	0.6 %	0.7 %
Value of production on Total Assets	0.3 %	0.2 %	0.7 %
Annual average of credit movements	0.3 %	0.4 %	0.7 %
Short-term payables on Net worth	0.3 %	0.2 %	0.7 %
Annual average of debit movements	0.3 %	0.3 %	0.3 %
Leverage	0.3 %	0.3 %	0.3 %
Debt to banks	0.2 %	0.3 %	0.3 %
Inventory turnover	0.2 %	0.3 %	0.3 %
Short-term debt	0.2 %	0.2 %	0.3 %
Acid test	0.2 %	0.3 %	0.7 %
Shareholders' equity on (long-term equity and payables)	0.2 %	0.3 %	0.3 %
Unit labor cost	0.2 %	0.1 %	0.3 %
Financial autonomy	0.2 %	0.2 %	0.3 %
Capital invested rotation	0.1 %	0.0 %	0.3 %
ROD	0.1 %	0.1 %	0.3 %

Table A5
XGBoost importance scores for sector 41.

	Gain	Cover	Frequency
Annual used credit on granted credit	11.0 %	7.2 %	5.0 %
Number of months of overdraft every six months	7.7 %	4.9 %	1.7 %
Annual number of months of overdraft	4.6 %	2.3 %	1.2 %
Annual total credit/debit amounts on revenues	4.5 %	3.2 %	2.9 %
Annual used credit on granted credit	4.2 %	5.4 %	3.3 %
Inventories on Bank debt	2.0 %	2.4 %	2.0 %
Annual average of debit movements	1.9 %	2.7 %	2.4 %
Tangible assets on Total assets	1.8 %	2.4 %	2.4 %
ROE	1.7 %	2.5 %	2.1 %
ROA	1.7 %	2.2 %	2.0 %
Depreciation and devaluation on costs	1.7 %	1.2 %	1.5 %
Index of rigidity of assets	1.6 %	1.5 %	2.0 %
Financial dependence index	1.6 %	2.4 %	2.3 %
Annual average of credit movements	1.6 %	1.7 %	2.3 %
Annual total credit amounts on debit amounts	1.6 %	1.8 %	2.1 %
Immediate liquidity on total assets	1.5 %	1.4 %	2.0 %
Degree of indebtedness	1.5 %	1.1 %	2.0 %
Percentage of debit movements out of total number of annual debit/credit movements	1.5 %	1.4 %	1.7 %
Debt to banks	1.5 %	1.3 %	1.8 %
Fixed asset coverage	1.5 %	2.3 %	2.1 %
Inventories on Total debt	1.4 %	1.7 %	1.7 %
Annual overdraft on granted credit (compensated)	1.4 %	1.0 %	1.7 %
Payables to suppliers on Net worth (shareholders' equity)	1.4 %	1.6 %	1.8 %
Annual average of credit/debit movements	1.4 %	1.2 %	1.8 %
Annual overdraft on granted credit	1.3 %	1.4 %	1.5 %
Added value on production value	1.3 %	1.7 %	1.7 %
ROI	1.3 %	1.0 %	1.2 %
Inventory turnover	1.3 %	1.2 %	1.8 %
Annual total credit movements on debit movements	1.2 %	1.5 %	1.8 %
Value of production on Total Assets	1.2 %	1.4 %	1.7 %
Production value on Inventory	1.2 %	1.6 %	1.7 %
Short-term payables on Net worth	1.2 %	1.2 %	1.5 %
Payables to suppliers on Total debt	1.2 %	2.1 %	1.5 %
Working capital turnover	1.2 %	0.9 %	1.4 %
Short-term payables on amounts due to banks	1.1 %	1.4 %	1.5 %
Credits on Total assets	1.1 %	1.6 %	1.7 %

(*continued on next page*)

Table A5 (*continued*)

	Gain	Cover	Frequency
Gross operating profitability (EBITDA on production value)	1.1 %	0.9 %	1.5 %
Financial coverage index	1.0 %	1.1 %	1.4 %
medium and long term debt indebtness	1.0 %	1.2 %	1.5 %
Immediate liquidity	1.0 %	1.0 %	1.4 %
Tangible assets on shareholders' equity	0.9 %	1.2 %	1.1 %
Financial autonomy	0.9 %	1.2 %	0.9 %
Highest value of unpaid checks per year	0.9 %	1.2 %	0.8 %
Added value on Revenues	0.9 %	1.1 %	0.9 %
Current assets / current liabilities	0.9 %	1.3 %	1.4 %
Continuous months of overdraft per year	0.9 %	0.9 %	0.6 %
Short-term debt	0.9 %	0.7 %	1.1 %
Continuous months of overdraft every six months	0.9 %	0.6 %	0.6 %
Payables on short-term debts	0.8 %	1.2 %	1.1 %
Annual total credit amounts on debit/credit amounts	0.8 %	0.7 %	1.5 %
ROS	0.8 %	1.1 %	1.2 %
Shareholders' equity on (long-term equity and payables)	0.7 %	0.4 %	1.2 %
Other revenues on production value	0.7 %	0.6 %	1.2 %
Leverage	0.7 %	0.9 %	0.9 %
Inventories on Short-term debt	0.7 %	0.5 %	0.9 %
Capital invested rotation	0.7 %	1.0 %	1.2 %
Payables to banks on current assets	0.6 %	0.5 %	0.9 %
Average value of checks unpaid at first presentation per year	0.6 %	1.2 %	0.8 %
Inventory duration	0.5 %	0.2 %	0.6 %
EBITDA on revenues	0.5 %	0.3 %	0.8 %
Number of months of overdraft (compensated) every six months	0.5 %	0.8 %	0.3 %
Debts on Net worth	0.4 %	0.7 %	0.5 %
Annual continuous months of overdraft	0.4 %	0.3 %	0.5 %
Continuous months of overdraft (compensated) every six months	0.4 %	0.3 %	0.5 %
Inventories on Total assets	0.3 %	0.6 %	0.6 %
Total of checks unpaid at first presentation per year	0.3 %	0.7 %	0.6 %
Highest value of checks unpaid at first presentation per year	0.3 %	0.3 %	0.3 %
cc_flag12_s	0.3 %	0.3 %	0.2 %
cc_flag6_sc	0.3 %	0.3 %	0.2 %
Net active coverage	0.2 %	0.5 %	0.5 %
Inventories on medium and long term debt	0.1 %	0.1 %	0.2 %
Average value of unpaid checks per year	0.1 %	0.1 %	0.2 %
Annual number of months of overdraft (compensated)	0.1 %	0.3 %	0.2 %
Shareholders' equity on equity and inventories	0.0 %	0.0 %	0.2 %

Table A6
XGBoost importance scores for sector 43.

	Gain	Cover	Frequency
Annual used credit on granted credit	8.8 %	7.5 %	3.0 %
Annual total credit/debit amounts on revenues	6.7 %	5.8 %	4.2 %
Annual overdraft on granted credit (compensated)	3.3 %	3.0 %	1.1 %
Annual overdraft on granted credit	2.8 %	2.8 %	1.7 %
Annual used credit on granted credit	2.8 %	4.3 %	2.6 %
Number of months of overdraft (compensated) every six months	2.3 %	1.5 %	0.6 %
Degree of indebtedness	2.2 %	4.0 %	1.6 %
Labour cost on revenues	2.2 %	2.2 %	2.5 %
Immediate liquidity on total assets	2.1 %	1.7 %	2.8 %
Short-term payables on amounts due to banks	1.8 %	1.9 %	1.7 %
Payables to banks on current assets	1.6 %	1.0 %	1.9 %
Debt to banks	1.6 %	1.1 %	2.0 %
Highest value of checks unpaid at first presentation per year	1.5 %	3.4 %	1.2 %
Annual average of credit/debit movements	1.5 %	1.1 %	1.5 %
Annual total credit movements on debit movements	1.5 %	1.9 %	2.4 %
Other revenues on production value	1.5 %	1.4 %	2.2 %
Inventories on Short-term debt	1.5 %	1.5 %	1.8 %
Annual average of debit movements	1.4 %	1.6 %	2.0 %
Average value of checks unpaid at first presentation per year	1.4 %	1.1 %	1.1 %
Labor cost on production value	1.4 %	1.4 %	1.4 %
Index of rigidity of assets	1.3 %	1.2 %	1.5 %
Number of months of overdraft every six months	1.3 %	1.2 %	0.5 %
Payables to suppliers on Total debt	1.3 %	1.7 %	2.0 %
Added value on Revenues	1.3 %	1.3 %	1.5 %

(*continued on next page*)

Table A6 (*continued*)

	Gain	Cover	Frequency
Continuous months of overdraft per year	1.3 %	1.7 %	0.6 %
Inventories on Total assets	1.3 %	1.4 %	1.4 %
Depreciation and devaluation on costs	1.3 %	1.3 %	1.8 %
Payables to suppliers on Net worth (shareholders' equity)	1.3 %	0.9 %	1.6 %
Working capital turnover	1.3 %	1.6 %	1.7 %
Intangible assets on Total assets	1.3 %	0.8 %	1.1 %
Unit labor cost	1.2 %	1.7 %	1.4 %
Financial dependence index	1.2 %	0.9 %	1.6 %
Short-term debt	1.2 %	1.3 %	1.4 %
Shareholders' equity on equity and inventories	1.2 %	1.6 %	1.2 %
Intangible assets on shareholders' equity	1.2 %	0.9 %	1.1 %
Leverage	1.2 %	0.6 %	0.8 %
Acid test	1.1 %	0.6 %	1.2 %
Shareholders' equity on (long-term equity and payables)	1.1 %	0.9 %	0.9 %
ROA	1.1 %	0.8 %	2.0 %
Value of production on Total Assets	1.0 %	1.1 %	1.7 %
Short-term payables on Net worth	1.0 %	1.0 %	1.1 %
Credits on Total assets	1.0 %	1.0 %	1.7 %
Unit cash flow (on total revenues)	1.0 %	0.7 %	1.4 %
ROE	1.0 %	1.1 %	1.5 %
Annual average of credit movements	1.0 %	0.9 %	1.1 %
Annual total credit amounts on debit/credit amounts	0.9 %	0.6 %	1.0 %
Financial coverage index	0.9 %	1.2 %	1.3 %
Annual total credit amounts on debit amounts	0.9 %	1.1 %	1.5 %
Fixed asset coverage	0.9 %	0.6 %	1.0 %
Annual number of months of overdraft	0.8 %	0.6 %	0.6 %
Debts on Net worth	0.7 %	0.7 %	0.6 %
Inventory duration	0.7 %	0.5 %	0.8 %
Tangible assets on shareholders' equity	0.7 %	1.0 %	1.1 %
Total of checks unpaid at first presentation per year	0.7 %	1.0 %	0.6 %
ROS	0.7 %	0.7 %	1.3 %
Added value on production value	0.7 %	0.5 %	1.3 %
Capital invested rotation	0.7 %	0.9 %	1.0 %
Tangible assets on Total assets	0.7 %	0.6 %	0.9 %
Net active coverage	0.6 %	0.7 %	0.6 %
Immediate liquidity	0.6 %	0.4 %	1.3 %
Inventories on Total debt	0.6 %	0.6 %	0.6 %
ROI	0.6 %	0.6 %	1.2 %
Percentage of debit movements out of total number of annual debit/credit movements	0.6 %	0.6 %	0.7 %
medium and long term debt indebtness	0.5 %	0.2 %	0.6 %
Gross operating profitability (EBITDA on production value)	0.5 %	0.5 %	0.7 %
EBITDA on revenues	0.5 %	0.2 %	0.4 %
Cash Flow on production value	0.5 %	0.4 %	1.0 %
Current assets / current liabilities	0.5 %	1.0 %	0.8 %
Annual continuous months of overdraft	0.5 %	0.3 %	0.3 %
Annual number of months of overdraft (compensated)	0.5 %	0.5 %	0.3 %
Inventories on Bank debt	0.5 %	0.5 %	0.9 %
Inventory turnover	0.4 %	0.4 %	0.6 %
Intangible asset on production value	0.4 %	0.5 %	0.6 %
Production value on Inventory	0.4 %	0.4 %	1.0 %
cc_flag6_s	0.4 %	1.0 %	0.1 %
Payables on short-term debts	0.4 %	0.2 %	0.5 %
Highest value of unpaid checks per year	0.3 %	1.0 %	0.3 %
Financial autonomy	0.3 %	0.1 %	0.2 %
Average value of unpaid checks per year	0.3 %	0.3 %	0.2 %
Value of unpaid checks on revenues	0.2 %	0.7 %	0.3 %
Total unpaid checks per year	0.1 %	0.0 %	0.1 %
Inventories on medium and long term debt	0.1 %	0.0 %	0.1 %
Continuous months of overdraft every six months	0.1 %	0.2 %	0.2 %

Table A7
XGBoost importance scores for sector 45.

	Gain	Cover	Frequency
Annual used credit on granted credit	5.8 %	6.6 %	4.3 %
Annual total credit/debit amounts on revenues	4.6 %	2.6 %	3.3 %
Annual used credit on granted credit	4.0 %	3.8%	2.7%

(*continued on next page*)

Table A7 (*continued*)

	Gain	Cover	Frequency
Annual average of credit/debit movements	3.3%	2.0%	2.3%
Degree of indebtedness	2.8%	1.5%	2.3%
Immediate liquidity on total assets	2.6%	2.9%	3.0%
ROA	2.4%	1.9%	2.3%
Annual overdraft on granted credit (compensated)	2.4%	5.6%	1.7%
Short-term payables on Net worth	2.2%	1.0%	1.7%
Average value of checks unpaid at first presentation per year	2.2%	4.6%	0.7%
Annual average of debit movements	2.2%	0.8%	2.0%
Percentage of debit movements out of total number of annual debit/credit movements	2.1%	1.6%	2.0%
Annual overdraft on granted credit	2.0%	1.9%	1.7%
Intangible asset on production value	2.0%	1.4%	1.3%
Annual total credit amounts on debit amounts	1.9%	1.3%	2.0%
Unit labor cost	1.9%	1.8%	1.7%
Continuous months of overdraft every six months	1.9%	1.7%	0.3%
Immediate liquidity	1.7%	1.1%	1.7%
Annual total credit movements on debit movements	1.7%	0.9%	3.0%
Short-term payables on amounts due to banks	1.7%	0.9%	1.3%
Annual average of credit movements	1.6%	1.6%	2.3%
Total of checks unpaid at first presentation per year	1.6%	2.0%	1.3%
Payables to suppliers on Total debt	1.6%	1.1%	2.0%
Other revenues on production value	1.6%	1.6%	2.0%
Financial interest on revenues	1.5%	2.5%	1.3%
ROE	1.5%	0.7%	1.7%
Intangible assets on shareholders’ equity	1.4%	1.6%	1.7%
Tangible assets on Total assets	1.4%	0.9%	1.7%
Depreciation and devaluation on costs	1.3%	1.1%	1.7%
Capital invested rotation	1.3%	0.3%	1.0%
Unit cash flow (on total revenues)	1.3%	0.9%	1.3%
ROS	1.3%	3.2%	1.3%
Highest value of unpaid checks per year	1.2%	0.5%	0.7%
Index of rigidity of assets	1.2%	0.6%	1.0%
Fixed asset coverage	1.2%	0.6%	2.0%
Inventory turnover	1.2%	0.8%	1.3%
Inventories on Total assets	1.2%	1.5%	1.3%
Annual number of months of overdraft	1.2%	2.8%	0.3%
Financial dependence index	1.2%	0.5%	1.3%
Inventories on Short-term debt	1.1%	0.8%	1.7%
Working capital turnover	1.1%	1.0%	1.7%
Cash Flow on production value	1.0%	2.7%	1.0%
Gross operating profitability (EBITDA on production value)	1.0%	0.8%	1.3%
Annual total credit amounts on debit/credit amounts	1.0%	0.7%	1.3%
Added value on production value	1.0%	1.0%	1.7%
Value of production on Total Assets	0.9%	0.6%	0.7%
Payables to banks on current assets	0.9%	0.4%	1.0%
Current assets / current liabilities	0.8%	0.5%	1.0%
Financial coverage index	0.8%	0.4%	1.0%
Shareholders’ equity on equity and inventories	0.8%	2.2%	1.3%
Inventory duration	0.7%	0.4%	0.7%
Financial autonomy	0.7%	0.3%	1.0%
Inventories on Bank debt	0.7%	0.5%	0.7%
Labour cost on revenues	0.7%	0.5%	0.7%
Tangible assets on shareholders’ equity	0.7%	0.4%	0.7%
Labor cost on production value	0.7%	0.6%	1.0%
Inventories on Total debt	0.6%	0.5%	0.3%
Highest value of checks unpaid at first presentation per year	0.6%	0.8%	0.3%
Added value on Revenues	0.6%	0.4%	1.0%
cc_flag6_s	0.6%	0.6%	0.3%
Debt burden index	0.6%	0.8%	0.3%
Financial interest on Added value	0.6%	5.7%	1.0%
Inventories on medium and long term debt	0.4%	0.2%	1.0%
Acid test	0.4%	0.4%	0.3%
Debt to banks	0.4%	0.2%	0.3%
Intangible assets on Total assets	0.4%	0.2%	0.7%
Production value on Inventory	0.4%	0.4%	1.0%
Payables on short-term debts	0.3%	0.3%	1.0%
Continuous months of overdraft per year	0.3%	0.5%	0.7%
Debts on Net worth	0.3%	0.4%	0.3%
Credits on Total assets	0.3%	0.2%	0.3%
Net active coverage	0.3%	0.3%	0.7%
Payables to suppliers on Net worth (shareholders’ equity)	0.3%	0.3%	1.0%

(*continued on next page*)

Table A7 (*continued*)

	Gain	Cover	Frequency
Short-term debt	0.2%	0.3%	0.7%
Average value of unpaid checks per year	0.2%	1.6%	0.3%
Number of months of overdraft every six months	0.1%	0.3%	0.3%
medium and long term debt indebtness	0.1%	0.0%	0.3%
ROD	0.1%	1.0%	0.3%

Table A8
XGBoost importance scores for sector 46.

	Gain	Cover	Frequency
Annual used credit on granted credit	5.5%	9.6%	3.1%
Annual overdraft on granted credit	4.9%	5.8%	2.5%
Annual total credit/debit amounts on revenues	4.8%	3.5%	4.1%
Annual number of months of overdraft (compensated)	4.6%	3.9%	0.5%
Annual used credit on granted credit	4.1%	8.6%	4.2%
Number of months of overdraft every six months	3.7%	3.9%	0.8%
Payables to suppliers on Total debt	2.6%	2.5%	2.7%
Annual average of debit movements	2.5%	2.4%	3.0%
Annual number of months of overdraft	2.4%	2.1%	0.5%
Other revenues on production value	2.3%	2.0%	2.7%
Annual total credit amounts on debit amounts	2.2%	1.3%	2.5%
Annual total credit movements on debit movements	2.1%	1.3%	2.2%
Depreciation and devaluation on costs	2.0%	1.8%	2.2%
Working capital turnover	2.0%	1.2%	2.2%
Degree of indebtedness	2.0%	1.7%	1.4%
ROE	1.9%	2.7%	2.2%
ROA	1.9%	3.3%	2.8%
Added value on Revenues	1.6%	1.6%	2.0%
Index of rigidity of assets	1.6%	1.1%	1.6%
Immediate liquidity	1.6%	0.9%	1.9%
Fixed asset coverage	1.6%	2.2%	2.2%
Intangible assets on Total assets	1.5%	1.8%	1.9%
Payables to banks on current assets	1.5%	1.2%	1.9%
Labour cost on revenues	1.4%	0.6%	1.4%
Debt to banks	1.4%	0.5%	1.6%
Payables to suppliers on Net worth (shareholders' equity)	1.3%	0.8%	1.6%
Short-term debt	1.2%	0.5%	1.3%
Labor cost on production value	1.2%	1.2%	1.3%
Value of production on Total Assets	1.2%	0.6%	1.7%
Unit cash flow (on total revenues)	1.2%	0.6%	1.4%
Current assets / current liabilities	1.1%	0.8%	1.3%
Credits on Total assets	1.1%	1.2%	1.9%
Inventories on Bank debt	1.1%	1.0%	1.6%
ROI	1.1%	0.6%	1.3%
Inventories on Total debt	1.1%	1.7%	1.4%
Debts on Net worth	1.1%	1.0%	1.3%
Inventories on Total assets	1.1%	0.8%	1.1%
medium and long term debt indebtness	1.0%	0.5%	1.1%
Percentage of debit movements out of total number of annual debit/credit movements	1.0%	0.8%	1.1%
Tangible assets on shareholders' equity	1.0%	1.0%	1.3%
Inventories on Short-term debt	0.9%	0.9%	1.1%
Unit labor cost	0.9%	0.8%	1.3%
Highest value of checks unpaid at first presentation per year	0.9%	1.1%	1.1%
Immediate liquidity on total assets	0.9%	0.7%	1.3%
Annual average of credit/debit movements	0.9%	0.8%	1.1%
Production value on Inventory	0.9%	1.2%	1.3%
Average value of unpaid checks per year	0.8%	0.6%	0.5%
Capital invested rotation	0.8%	0.8%	1.1%
Cash Flow on production value	0.8%	1.0%	0.9%
Tangible assets on Total assets	0.7%	0.6%	1.3%
Short-term payables on amounts due to banks	0.7%	0.3%	0.9%
Short-term payables on Net worth	0.7%	0.9%	0.8%
EBITDA on revenues	0.7%	0.5%	0.8%
Annual average of credit movements	0.7%	0.5%	0.9%
Intangible asset on production value	0.6%	0.6%	0.8%
Annual total credit amounts on debit/credit amounts	0.6%	0.1%	0.8%
ROS	0.6%	0.9%	0.9%

(*continued on next page*)

Table A8 (*continued*)

	Gain	Cover	Frequency
Financial autonomy	0.6%	0.2%	0.9%
Shareholders' equity on (long-term equity and payables)	0.6%	0.7%	0.8%
Added value on production value	0.6%	0.4%	0.8%
Gross operating profitability (EBITDA on production value)	0.5%	0.4%	0.5%
Inventory duration	0.5%	1.1%	0.8%
Financial dependence index	0.5%	0.2%	0.9%
Continuous months of overdraft every six months	0.5%	0.7%	0.5%
Payables on short-term debts	0.4%	0.6%	0.6%
Annual overdraft on granted credit (compensated)	0.3%	0.4%	0.5%
Inventory turnover	0.3%	0.2%	0.6%
Number of months of overdraft (compensated) every six months	0.3%	0.1%	0.3%
Total of checks unpaid at first presentation per year	0.2%	0.5%	0.3%
Annual continuous months of overdraft	0.2%	0.3%	0.6%
Net active coverage	0.2%	0.0%	0.2%
Average value of checks unpaid at first presentation per year	0.2%	1.1%	0.3%
Shareholders' equity on equity and inventories	0.2%	0.1%	0.3%
Highest value of unpaid checks per year	0.2%	0.1%	0.2%
Leverage	0.1%	0.0%	0.2%

Table A9
XGBoost importance scores for sector 47.

	Gain	Cover	Frequency
Annual overdraft on granted credit	4.8%	7.1%	3.0%
Annual total credit/debit amounts on revenues	4.1%	3.1%	3.5%
Annual used credit on granted credit	3.5%	3.9%	2.9%
Annual used credit on granted credit	2.9%	3.8%	2.4%
Number of months of overdraft every six months	2.4%	2.1%	0.8%
ROA	2.2%	2.8%	1.9%
Value of production on Total Assets	2.2%	1.9%	1.4%
Immediate liquidity on total assets	2.1%	2.4%	2.3%
Inventories on Short-term debt	2.1%	1.5%	1.9%
Annual average of credit movements	2.1%	1.9%	2.4%
ROE	2.0%	1.7%	2.7%
Annual total credit movements on debit movements	2.0%	1.7%	2.2%
Annual number of months of overdraft (compensated)	1.9%	2.3%	0.3%
Other revenues on production value	1.9%	1.5%	2.1%
Intangible asset on production value	1.8%	1.3%	1.8%
Short-term debt	1.8%	2.0%	1.4%
Annual total credit amounts on debit amounts	1.8%	1.1%	2.0%
Financial dependence index	1.7%	1.6%	1.6%
Continuous months of overdraft every six months	1.7%	1.7%	0.4%
Annual continuous months of overdraft	1.6%	1.3%	0.6%
Financial coverage index	1.6%	1.8%	1.9%
Intangible assets on Total assets	1.6%	1.7%	1.9%
Depreciation and devaluation on costs	1.6%	1.2%	1.8%
Immediate liquidity	1.5%	1.6%	1.4%
Working capital turnover	1.5%	1.9%	1.9%
Debt to banks	1.5%	1.6%	1.8%
Inventories on Total assets	1.4%	0.8%	1.6%
Payables to suppliers on Total debt	1.4%	1.1%	2.1%
Current assets / current liabilities	1.4%	1.5%	1.9%
Tangible assets on shareholders' equity	1.4%	1.3%	1.5%
ROI	1.4%	1.3%	2.0%
Annual average of debit movements	1.3%	1.0%	1.3%
Shareholders' equity on equity and inventories	1.3%	1.0%	1.8%
Labor cost on production value	1.2%	0.8%	1.3%
Added value on Revenues	1.2%	1.0%	1.1%
Inventories on Total debt	1.2%	1.3%	1.1%
Payables to suppliers on Net worth (shareholders' equity)	1.2%	1.1%	1.1%
Index of rigidity of assets	1.1%	0.8%	1.5%
Labour cost on revenues	1.1%	0.9%	1.6%
Intangible assets on shareholders' equity	1.1%	1.2%	1.4%
medium and long term debt indebtness	1.1%	1.3%	1.4%
Credits on Total assets	1.1%	0.9%	1.1%
Annual number of months of overdraft	1.0%	1.0%	0.7%
Degree of indebtedness	1.0%	1.3%	1.4%

(*continued on next page*)

Table A9 (*continued*)

	Gain	Cover	Frequency
Unit cash flow (on total revenues)	1.0%	0.6%	1.1%
ROS	1.0%	0.9%	1.0%
Payables on short-term debts	1.0%	0.6%	0.9%
Cash Flow on production value	1.0%	0.9%	1.1%
Annual average of credit/debit movements	0.9%	0.5%	0.9%
Unit labor cost	0.9%	0.7%	1.0%
Added value on production value	0.9%	1.1%	1.1%
EBITDA on revenues	0.9%	0.7%	1.0%
Inventories on Bank debt	0.9%	0.8%	1.2%
Production value on Inventory	0.8%	0.6%	0.9%
Gross operating profitability (EBITDA on production value)	0.8%	0.8%	0.8%
Average value of unpaid checks per year	0.8%	1.9%	0.3%
Highest value of unpaid checks per year	0.7%	1.5%	0.7%
Inventory duration	0.7%	0.7%	1.0%
Short-term payables on amounts due to banks	0.7%	1.1%	1.1%
Shareholders' equity on (long-term equity and payables)	0.7%	0.9%	1.0%
Average value of checks unpaid at first presentation per year	0.7%	0.6%	0.9%
Tangible assets on Total assets	0.7%	0.5%	0.8%
Annual overdraft on granted credit (compensated)	0.7%	1.5%	0.8%
Capital invested rotation	0.7%	0.5%	0.9%
Total of checks unpaid at first presentation per year	0.7%	1.6%	0.7%
Fixed asset coverage	0.7%	0.7%	0.5%
Short-term payables on Net worth	0.6%	0.5%	0.8%
Value of unpaid checks on revenues	0.5%	0.6%	0.5%
Continuous months of overdraft per year	0.4%	0.2%	0.3%
Net active coverage	0.4%	0.4%	0.4%
Highest value of checks unpaid at first presentation per year	0.4%	0.5%	0.5%
Payables to banks on current assets	0.3%	0.4%	0.5%
Leverage	0.3%	0.3%	0.4%
Continuous months of overdraft (compensated) every six months	0.3%	0.1%	0.4%
Percentage of debit movements out of total number of annual debit/credit movements	0.3%	0.3%	0.3%
Debts on Net worth	0.2%	0.2%	0.3%
Inventory turnover	0.2%	0.2%	0.2%
Number of months of overdraft (compensated) every six months	0.2%	0.1%	0.3%
Annual total credit amounts on debit/credit amounts	0.2%	0.3%	0.3%
Total unpaid checks per year	0.1%	0.1%	0.1%
Financial autonomy	0.1%	0.0%	0.1%

Table A10
Stepwise logistic regression and XgBoost confusion matrices for sector 25, cutoff 0.2.

Stepwise Logistic		Actual		Total		XgBoost		Actual	Total
		default	no default		default			no default	
Fitted	default	14	42	56	Fitted	default	14	40	54
	no default	40	1136	1176		no default	40	1138	1178
	Total	54	1178	1232		Total	54	1178	1232

Table A11
Stepwise logistic regression and XgBoost confusion matrices for sector 28, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	4	30	34	Fitted	default	6	16	22
	no default	12	559	571		no default	10	573	583
	Total	16	589	605		Total	16	589	605

Table A12
Stepwise logistic regression and XgBoost confusion matrices for sector 41, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	131	239	370	Fitted	default	115	193	308
	no default	63	1103	1166		no default	79	1149	1228
	Total	194	1342	1536		Total	194	1342	1536

Table A13
Stepwise logistic regression and XgBoost confusion matrices for sector 43, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	37	77	114	Fitted	default	38	50	88
	no default	47	988	1035		no default	46	1015	1061
	Total	84	1065	1149		Total	84	1065	1149

Table A14
Stepwise logistic regression and XgBoost confusion matrices for sector 45, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	6	38	44	Fitted	default	8	41	49
	no default	14	564	578		no default	12	561	573
	Total	20	602	622		Total	20	602	622

Table A15
Stepwise logistic regression and XgBoost confusion matrices for sector 46, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	40	95	135	Fitted	default	49	108	157
	no default	89	2412	2501		no default	80	2399	2479
	Total	129	2507	2636		Total	129	2507	2636

Table A16
Stepwise logistic regression and XgBoost confusion matrices for sector 47, cutoff 0.2.

Stepwise Logistic		Actual		Total	XgBoost		Actual		Total
		default	no default				default	no default	
Fitted	default	26	65	91	Fitted	default	22	65	87
	no default	42	1014	1056		no default	46	1014	1060
	Total	68	1079	1147		Total	68	1079	1147

Table A17
Stepwise logistic regression forecast indexes, cutoff 0.2.

	25	28	41	43	45	46	47
Sensitivity	25.93%	25.00%	67.53%	44.05%	30.00%	31.01%	38.24%
Specificity	96.43%	94.91%	82.19%	92.77%	93.69%	96.21%	93.98%
Positive predictive value	25.00%	11.76%	35.41%	32.46%	13.64%	29.63%	28.57%
Negative predictive value	96.60%	97.90%	94.60%	95.46%	97.58%	96.44%	96.02%
Correctly classified	93.34%	93.06%	80.34%	89.21%	91.64%	93.02%	90.67%

Table A18
XGBoost forecast indexes, cutoff 0.2.

	25	28	41	43	45	46	47			
Sensitivity	25.93%	37.50%	59.28%	45.24%	40.00%	37.98%	32.35%			
Specificity	96.60%	97.28%	85.62%	95.31%	93.19%	95.69%	93.98%			
Positive predictive value	25.93%	27.27%	37.34%	43.18%	16.33%	31.21%	25.29%			
Negative predictive value	96.60%	98.28%	93.57%	95.66%	97.91%	96.77%	95.66%			
Stepwise Logistic		Actual	Total		XgBoost		Actual	Total		
		default	no default					default	no default	
Fitted	default	6	38	44		Fitted	default	8	41	49
	no default	14	564	578			no default	12	561	573
	Total	20	602	622			Total	20	602	622
Correctly classified	93.51%	95.70%	82.29%	91.64%	91.48%	92.87%	90.32%			

References

Altman, E.I., Esentato, M., Sabato, G., 2020. Assessing the credit worthiness of Italian SMEs and minibond issuers. Glob. Financ. J.

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. 23, 589–609.

Baesens, B., Van Gestel, T., 2009. Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital. Oxford University Press, Oxford.

Banasik, J., Crook, J., Thomas, L., 2003. Sample selection bias in credit scoring models. J. Oper. Res. Soc. vol. 54 (8), 822–832.

Beaver, W.H., 1966. Financial ratios as predictors of failure. J. Account. Res. 4, 71–111.

Blochlinger, A., Leippold, M., 2006. Economic benefit of powerful credit scoring. J. Bank. Financ. 30 (3), 851–873.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794.

Doumpos, M., Lemonakis, C., Niklis, D., Zopounidis, C., 2019. Analytical Techniques in the Assessment of Credit Risk. Springer, New York.

Durand, D., 1941. Risk Elements in Consumer Instalment Financing. NBER Books.

Filipe, S.F., Grammatikos, T., Michala, D., 2016. Forecasting distress in European SME portfolios. J. Bank. Financ. *64*, 112–135.

Galli, E., Mascia, D.V., Rossi, S.P.S., 2018. Does corruption influence the self-restraint attitude of women-led SMEs towards bank lending? CESifo Econ. Stud. 64 (3), 426–455.

Galli, E., Mascia, D.V., Rossi, S.P.S., 2020. Bank credit constraints for women-led SMEs: self-restraint or lender bias? Eur. Financ. Manag. 26 (4), 1147–1188.

Gunnarsson, B.R., vanden Broucke, S., Baesens, B., Óskarsdóttir, M., Lemahieu, W., 2021. Deep learning for credit scoring: do or don't? Eur. J. Oper. Res. 295, 292–305.

Gupta, J., Gregoriou, A., Healy, J., 2015. Forecasting bankruptcy for SMEs using hazard function: to what extent does size matter? Rev. Quant. Financ. Account. 45 (4), 845–869.

Hajek, P., Abedin, M.Z., Sivarajah, U., 2023. Fraud detection in mobile payment systems using an XGBoost-based framework. Inf. Syst. Front. 25 (5), 1985–2003.

He, H., Zhang, W., Zhang, S., 2018. A novel ensemble method for credit scoring: adaption of different imbalance ratios. Expert Syst. Appl. 98, 105–117.

Lessmann, S., Baesens, B., Seow, H.-V., Thomas, L.C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur. J. Oper. Res. 247 (1), 124–136.

Llewellyn, D.T., 2012. The Evolution of Bank Business Models: Pre- and Post-crisis. in Future Risks and Fragilities for Financial Stability. SUERF Study 2012/3, Wien.

Liu, Yi, 2023. Design of XGBoost prediction model for financial operation fraud of listed companies. Int. J. Syst. Assur. Eng. Manag. Volume 14 (Issue 6), 2354–2364, 2023.

Maehara, R., Benites, L., Talavera, A., Aybar-Flores, A., Muñoz, M., 2024. Predicting financial inclusion in Peru: application of machine learning algorithms. J. Risk Financ. Manag. 17 (1), 34, 2024.

Marques, A., Garcıa, V., Sanchez, S., 2013. A literature review on the application of evolutionary computing to credit scoring. J. Oper. Res. Soc. 64 (9), 1384–1399.

Mascia, D.V., 2018. Young Enterprises and Bank Credit Denials. ADBI Working Paper 844. Asian Development Bank Institute, Tokyo. ⟨https://www.adb.org/publications/young-enterprises-and-bank-credit-denials⟩ (Available).

Mascia, D.V., Rossi, S.P.S., 2017. Is there a gender effect on the cost of bank financing? J. Financ. Stab. 2017 (31), 136–153.

Nti, I.K., Somanathan, A.R., 2022. A Scalable RF-XGBoost framework for financial fraud mitigation. IEEE Trans. Comput. Soc. Syst. 2022.

Rikkers, F., Thibeault, A.E., 2011. Default prediction of small and medium-sized enterprises with industry effects. Int. J. Bank. Account. Financ. 3 (2–3), 207–231.

Xia, Y., Liu, C., Li, Y., Liu, N., 2017. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Syst. Appl. 78, 225–241.

Xueping, Z., Samuri, S.M., Adnan, M.H.M., 2023. Exploring the performance of XGBOOST and artificial neural network in personal credit default prediction: an empirical study. Int. Conf. Electr. Comput. Commun. Mechatron. Eng. ICECCME 2023.

ZHANG, Yuyan, CHEN, Ke, CHEN, Ting, 2023. Analysis and prediction of bank customer loyalty based on XGBoost Algorithm. Front. Artif. Intell. Appl. Volume 373, 631–640, 2023.

Zhiyao, T., Yiyi, H., Chi, J., Yin, Z., 2024. User financial credit analysis for blockchain regulation. Comput. Electr. Eng. 113.