BANCA D'ITALIA

EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

The Bank of Italy's statistical model for the credit assessment of non-financial firms

by Simone Narizzano, Marco Orlandi and Antonio Scalia

October 2024

Number 53

BANCA D'ITALIA
EUROSISTEMA

Mercati, infrastrutture, sistemi di pagamento

(Markets, Infrastructures, Payment Systems)

The Bank of Italy's statistical model for the credit assessment of non-financial firms

by Simone Narizzano, Marco Orlandi and Antonio Scalia

Number 53 – October 2024

The papers published in the 'Markets, Infrastructures, Payment Systems' series provide information and analysis on aspects regarding the institutional duties of the Bank of Italy in relation to the monitoring of financial markets and payment systems and the development and management of the corresponding infrastructures in order to foster a better understanding of these issues and stimulate discussion among institutions, economic actors and citizens.

The views expressed in the papers are those of the authors and do not necessarily reflect those of the Bank of Italy.

The series is available online at www.bancaditalia.it.

Printed copies can be requested from the Paolo Baffi Library: richieste.pubblicazioni@bancaditalia.it.

ISSN 2724-6418 (online)
ISSN 2724-640X (print)

Banca d'Italia
Via Nazionale, 91 - 00184 Rome - Italy
+39 06 47921

Designed and printing by the Printing and Publishing Division of the Bank of Italy

The Bank of Italy's statistical model for the credit assessment of non-financial firms

by Simone Narizzano, Marco Orlandi and Antonio Scalia*

Abstract

The Bank of Italy has been managing the in-house credit assessment system (ICAS) for Italian non-financial firms since 2013, a system used in the Eurosystem's collateral framework. The ICASes, also operating at other Eurosystem national central banks, play a crucial role in monetary policy implementation in the euro area as they allow all counterparties to pledge credit claims to non-financial firms, particularly during episodes of market distress. The Bank of Italy's ICAS rating process has two stages that combine the statistical model with an expert assessment, performed by two analysts and the rating committee, to obtain the final rating for the firm. Every month, the statistical model produces the probability of default (PD) over a one-year horizon for 370,000 non-financial firms, using a fully automated procedure. This paper illustrates the methodology underlying the Bank of Italy's ICAS statistical model and its validation process. The model preserves simplicity and 'readability' by relying on a logit regression, while it tries to improve predictive performance with machine learning components for some variables that display non-linear behaviour towards default prediction. The model shows robust properties, as it discriminates between healthy and risky firms with fairly stable results. The discriminatory power is rather high and it improves as the size of the company increases, thus ensuring a proper evaluation of the largest exposures in monetary policy operations.

JEL Classification: G32, G33, C51, C52.

Keywords: Credit Risk, Credit Scoring, Probability of Default, Collateral Framework.

Sintesi

Dal 2013 la Banca d'Italia gestisce il proprio sistema interno di valutazione del merito creditizio (ICAS) per le imprese non finanziarie italiane, utilizzato nel quadro delle garanzie dell'Eurosistema. Gli ICAS, sviluppati anche presso altre Banche centrali nazionali dell'Eurosistema, svolgono un ruolo cruciale nell'attuazione della politica monetaria nell'area dell'euro, poiché consentono a tutte le controparti di conferire a garanzia prestiti verso società non finanziarie, in particolare durante gli episodi di tensione sui mercati. Il processo di valutazione del merito di credito della Banca d'Italia prevede due stadi e combina il modello statistico con la valutazione esperta, effettuata da due analisti e dal comitato rating, per ottenere il rating finale dell'impresa. Il modello statistico produce ogni mese la probabilità di default (PD) a un anno per 370.000 imprese non finanziarie con una procedura automatizzata. In questo lavoro si illustra la metodologia del modello statistico ICAS della Banca d'Italia e il suo processo di validazione. Il modello è semplice e 'leggibile' poiché si basa su un approccio generale di tipo logit, mentre prova a migliorare la capacità predittiva mediante l'impiego di componenti di *machine learning* per alcune variabili che mostrano una relazione non

* Bank of Italy, Financial Risk Management Directorate.

lineare verso l'evento di insolvenza. Il modello mostra una struttura robusta e proprietà soddisfacenti che consentono di discriminare tra aziende sane e rischiose con una discreta stabilità dei risultati. Il potere discriminante è elevato e migliora con l'aumentare della dimensione delle imprese, garantendo così una corretta valutazione del rischio sulle maggiori esposizioni nelle operazioni di politica monetaria.

CONTENTS

1. Introduction[1]

The Bank of Italy has been managing the in-house credit assessment system (ICAS) for Italian non-financial firms since 2013, a system used in the Eurosystem's collateral framework. Similar systems are also operating at other Eurosystem national central banks (NCBs). The Bank of Italy's ICAS rating process is based on a two-stage procedure which combines a statistical model with an expert assessment, performed by two analysts and possibly the rating committee, to obtain the final rating of the firm. The 'full' ICAS rating, including the expert assessment, is regularly produced for 4,000 Italian firms that are debtors of a large share of the loans posted as collateral in refinancing operations. The statistical model, that is the first stage of the rating process, covers a much larger sample by means of a fully automated procedure. Every month the model produces estimates of the probability of default (PD) over a one-year horizon for about 370,000 non-financial firms, namely all those that: i) are incorporated as a limited liability company in Italy; ii) publish a financial statement; and iii) have an exposure towards the banking system, as reported in the National Credit Register (NCR).

While the credit assessment of firms is a well-established practice among banks and rating agencies, and the Bank of Italy's ICAS obviously shares many features with the models of private entities, it has one information advantage, because it takes advantage of full access to the detailed credit performance of each firm as recorded in the NCR. ICAS ratings, like those of commercial banks, are not public. The ICASes of the NCBs play a crucial role in monetary policy implementation in the euro area as they allow all counterparties to pledge credit claims to non-financial firms. The ICAS contribution to the transmission of monetary policy is even more important during episodes of market tension, as ICASes support the overall availability of collateral. This paper illustrates the methodology underlying the Bank of Italy's ICAS statistical model and its validation process.

Default forecasting is of key importance for financial institutions and investors. Banks use PDs to evaluate the risks stemming from their lending activity, while investors employ PDs for bond pricing and portfolio management. The related literature, developed since the 1960s, is substantial (see Altman, 1968; Edmister, 1972; Deakin, 1972; Blum, 1974). Although credit models vary widely, they can be divided into two main categories: statistical models and machine learning (ML) models. Several studies compare the performance of statistical models and ML models in corporate default prediction. The majority of these studies concludes that ML models on average slightly outperform statistical models thanks to the ability to capture variables with a non-linear or non-monotonic relationship with default. However, ML models are opaque for the credit analyst and this 'black box' feature makes their use and interpretation rather difficult (Partridge *et al.*, 2017). A recent strand of the literature has shown that hybrid credit scoring models, which integrate ML techniques with statistical methods, offer a promising approach to default prediction (Van Gestel *et al.*, 2005; Dumitrescu *et al.*, 2022). In the development of the statistical model we have followed the latter approach, which tries to solve the trade-off between predictability and transparency by combining features of traditional models and ML models. The ICAS statistical approach

[1] We are grateful to Aviram Levy, Francesco Columba, Filippo Giovannelli and Francesco Monterisi for their useful comments. We also thank all the colleagues at the Credit Risk Assessment Division and at the Financial Risk Control Division for their contribution in the development of the statistical model presented in this paper and in the validation procedure, respectively.

preserves simplicity and 'readability' by relying on a logit regression, while it tries to improve predictive performance with ML components for some variables that display a non-monotonic behaviour towards default prediction. For these variables we develop a discretized transformation obtained with a decision tree and employ the transformed variables in the logit regression.

Every month the model produces an estimate of the PD over a one-year horizon for virtually all Italian non-financial firms with a fully automated procedure. The model shows robust and satisfactory properties that enable a high discrimination between healthy and risky firms with a fair stability of the results. The discriminatory power is rather high and it improves as the size of the company increases, thus ensuring a proper evaluation of the largest exposures in monetary policy operations. The architecture of the model is rather flexible, enabling us to work on possible developments to further improve the model's discriminatory power.

The remainder of the paper is organized as follows. Section 2 outlines the Eurosystem framework within which the ICAS operates. Section 3 describes the architecture of the ICAS model. Section 4 reviews the main approaches used in the literature to estimate the probability of default of a firm. Section 5 provides details on the ICAS statistical model. Section 6 describes the validation process and section 7 presents the main results on performance and a comparison with the previous version of the model (Giovannelli *et al.*, 2020). Section 8 concludes.

2. The Eurosystem credit quality standards for collateral

The collateral framework is one of the key pillars that support the implementation of monetary policy. All Eurosystem liquidity providing operations require counterparties to provide adequate collateral. The concept of collateral adequacy has two dimensions. First, collateral has to protect the Eurosystem against losses in its credit operations, that might affect its financial independence and credibility. Second, there should be enough collateral potentially available, to enable the transmission of monetary policy and provide a level playing field for counterparties in each country in the area.

The eligibility requirements set by the Eurosystem are mainly aimed at mitigating financial, legal and operational risks incurred in the conduct of monetary policy. The Eurosystem Credit Assessment Framework (ECAF)[2] defines the minimum credit quality requirements as well as the rules and procedures which ensure that the Eurosystem only accepts adequate collateral, as required by article 18.1 of the Protocol on the Statute of the European System of Central Banks and of the European Central Bank.

With a view to accepting a very broad range of marketable and non-marketable assets as collateral, the Eurosystem relies on three sources of credit assessment:

- credit rating agencies accepted as external credit assessment institutions (ECAIs): five of them are currently in the list (DBRS, Fitch, Moody's, Scope Ratings, and Standard&Poor);[3]

[2] For more information, see https://www.ecb.europa.eu/paym/coll/risk/ecaf/html/index.en.html

[3] On 2 November 2023, the Governing Council decided to accept the credit rating agency Scope Ratings as a new ECAI for the purposes of the ECAF.

- in-house credit assessment systems (ICASes) managed by NCBs: such systems are currently operated in Austria, France, Germany, Greece, Ireland, Italy, Portugal, Slovenia, and Spain;
- internal rating-based (IRB) systems managed by banks: around 40 systems are currently authorized in the whole euro area.

ECAIs are mainly used for the assessment of marketable collateral, whereas ICASes and IRB systems are largely used for non-marketable assets. The ECAF maps each rating grade into a single harmonized rating scale to make the credit ratings comparable across systems and sources.

ECAF foresees that all accepted credit assessment systems are subject to due diligence and a yearly performance monitoring process. The monitoring process has two components: 1) quantitative statistical methods are employed to check whether each system has accurately predicted default rates and the mapping of the ratings is appropriate; 2) a qualitative assessment is performed on credit assessment processes and methodologies.

The ICASes of the NCBs play a key role in ECAF as they allow any counterparty to pledge credit claims to non-financial firms as collateral. In particular, ICASes can be used by small and medium-sized banks that do not have an IRB system and are not in a position to easily fund themselves through structured finance operations, such as asset-backed securities or covered bonds. The development of ICASes has thus contributed to increasing collateral availability for a wide range of counterparties with different business models, thus paving the way for a smooth implementation of monetary policy. The use of credit claims as collateral in monetary policy operations enables counterparties to use liquid assets to guarantee market financing operations and comply with regulatory liquidity requirements (Grandia *et al.*, 2019). Credit claims have a low opportunity cost as collateral, whereas marketable assets, such as sovereign bonds, are largely used as collateral in private repo transactions. ICASes provide an even greater contribution to the correct transmission of monetary policy in times of market tension, as they support the overall availability of collateral. This crucial role is confirmed by the measures adopted by the Eurosystem in April 2020, during the financial and economic crisis brought about by the pandemic, when the Governing Council decided to enlarge the scope of the so-called additional credit claim (ACC) framework, that allows NCBs to expand the eligibility rules for credit claims in their own jurisdiction. The effectiveness of these measures is shown by the significant increase in the use of ICAS assessment in several countries since 2020.

ICASes also offer other related benefits for the Eurosystem. First, they foster the direct transmission of monetary policy measures to the real economy, especially to small and medium-sized enterprises (SMEs) that do not issue marketable bond instruments. Second, supporting the pledge of credit claims contributes to diversifying risk in the Eurosystem balance sheet. Finally, the development of ICASes contributes to increasing the Eurosystem internal credit risk assessment capabilities as a complement to external ratings, thus reducing the reliance on ECAIs, as recommended by the Financial Stability Board (FSB, 2010).

3. The architecture of the In-house Credit Assessment System

3.1 Aims and governance

The Bank of Italy has been developing its own internal model for assessing the creditworthiness of Italian non-financial firms (Giovannelli *et al.*, 2023) since 2013, for use in the context of the Eurosystem's collateral framework. The credit ratings and other information on non-financial firms managed by ICAS are also employed for other purposes, such as financial stability analysis, economic research, and occasionally for banking supervision purposes.

ICAS is managed by the Directorate Financial Risk Management (D-FRM). Within D-FRM, the Credit Risk Assessment Division (CRA) is in charge of model development, production of ratings and coordination tasks, while the Financial Risk Control Division (FRC) is responsible for the validation of the model, thus ensuring separation of tasks.[4] Seven ICAS Divisions in the Bank of Italy's network of local branches cooperate with D-FRM in the rating production process.

3.2 Rating process

The rating process is based on a two-stage procedure, which combines the statistical model with the expert assessment by rating analysts to produce the final rating (Fig. 1).

Figure 1 – Rating process

The rating of a non-financial firm hinges on its estimated probability of default (PD) over a one-year horizon. In the first stage of the rating process the estimated PD is obtained from the statistical model; in the second stage the expert assessment is performed on the firm's credit worthiness. This process is similar to that in place for public ratings at the ECAIs.

[4] Among the authors of this paper, Antonio Scalia is the head of D-FRM, responsible for the whole ICAS process. Simone Narizzano and Marco Orlandi are members of the CRA Division and of the FRC Division, respectively.

The PD estimates are categorized into risk classes on the internal rating scale and the ratings are then mapped to the corresponding credit quality step (CQS) of the Eurosystem harmonized rating scale (Table 1).

Table 1 – Rating scale
(percentage values)

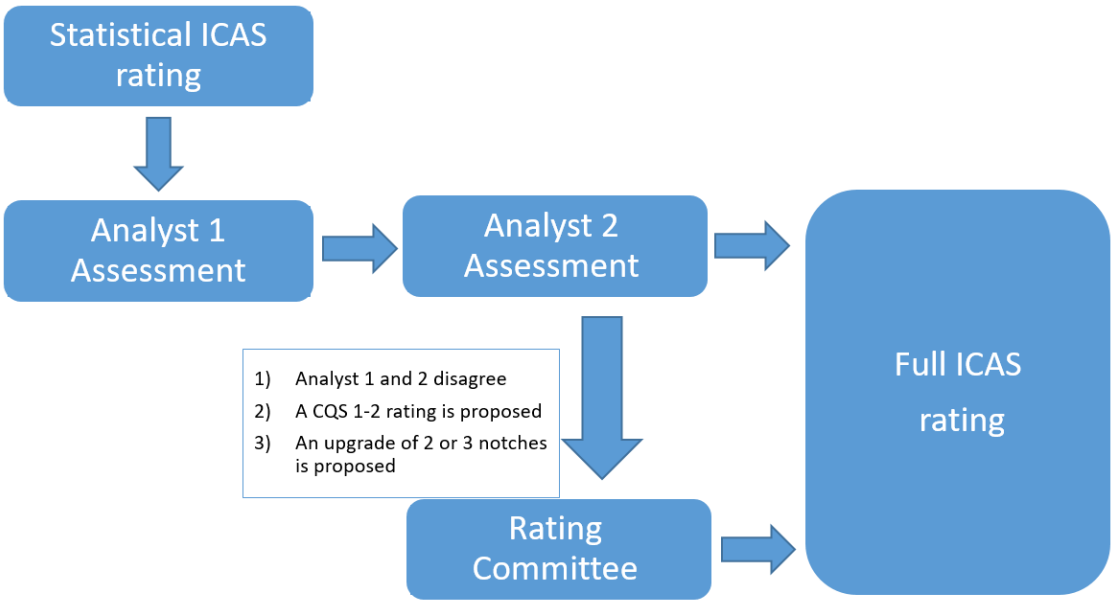
Risk Class of ICAS	Minimum PD	Maximum PD	Eurosystem Credit Quality Step
1	0.000	0.001	CQS 1 & 2
2+	0.001	0.01	
2	0.01	0.03	
2-	0.03	0.05	
3+	0.05	0.07	
3	0.07	0.09	
3-	0.09	0.10	
4+	0.10	0.17	CQS 3
4	0.17	0.30	
4-	0.30	0.40	
5+	0.40	0.80	CQS 4
5	0.80	1	
5-	1	1.50	CQS 5
6+	1.5	2	CQS 6
6	2	3	
6-	3	5	CQS 7
7	5	25	CQS8
8	25	100	
9	100	100	Default

Every month the statistical model yields the PD over a one-year horizon for about 370,000 Italian firms. The statistical model procedure is fully automated.

The expert assessment is performed by two credit analysts. Starting from the statistical PD, the first analyst reviews the main characteristics of the firm (size, business sector, geographical location, etc.), the statistical rating, and its components. Next, the analyst takes into account some potential risk profiles and determines a partial score for each of them. This reflects the opinion as to whether each profile improves, confirms or worsens the risk assessment obtained from the statistical model. The risk profiles include the following factors: financial ratios (peer group review); financial flexibility; quality of management, corporate governance and social awareness; economic environment, country risk, and industrial sector; group analysis; third-party opinions; climate risk (transition risk and physical risk); and recent news on the firm. Next, the grades for each profile are weighted and aggregated to obtain the final grade. The first

analyst can either confirm the rating obtained from the statistical model or revise the risk class up or down. The second analyst reviews all the previous steps and produces an independent assessment. The final assessment is upper-bounded: analysts can downgrade the rating without limitations, while they can raise the statistical rating by one notch only. If they intend to upgrade the statistical rating by more than one level, or if they disagree on the level, they are required to submit the proposal(s) to the ICAS rating committee for the final decision. The committee is composed of senior managers within D-FRM. The analysts involved in the assessment are also required to attend the committee meeting. The committee makes the final decision, which can involve a rating upgrade by up to three notches (Fig. 2).

Figure 2 – Rating decision process



4. Credit scoring: statistical models vs machine learning models

Default forecasting is of key importance for financial institutions and investors. Banks use PDs to screen potential borrowers, evaluate the terms of new loans, and manage the risks stemming from lending activities. Investors also make extensive use of PDs and probabilities of migration across different credit rating classes for bond pricing and portfolio management. Macroprudential authorities have an interest in the surveillance of default risk as it is a major source of risk for lenders.

Since the Basel II Accord, credit scoring methods have become popular in the banking industry. They apply multivariate models to the firm characteristics, such as economic and financial ratios, to predict its credit quality. The output of these methods is a credit score, namely a continuous numerical indicator of creditworthiness, which can be translated into a probability of default. Credit scoring models can be divided into two categories: statistical models and machine learning (ML) models.

Statistical models are particularly fit for the purpose of inference, since they rely on assumptions regarding the structural relationship between variables, the number of parameters that can be robustly estimated,

and the distribution properties of the data generating process. ML models are mostly aimed at prediction accuracy, and make very weak assumptions on the structure of the data generating process. This allows for the detection of data-driven interactions and non-linear or non-monotonic relationships between predictors and the outcome variable. A recent strand of the literature has shown that hybrid credit scoring models, which integrate ML techniques with traditional statistical methods, offer a promising approach to default prediction, balancing interpretability and discriminatory power.

In the remainder of this section we briefly describe these categories in turn.

4.1 Statistical models

The literature about default prediction with statistical methods is very wide. The pioneering work is Altman (1968), which develops univariate and multivariate methods to predict the firm bankruptcy using a set of financial ratios. Altman uses a multivariate linear discriminant analysis (LDA) to estimate the so-called Z-score model on a sample of manufacturing firms during the period 1946-1965. The study starts from a list of 22 financial ratios and ends up selecting five of them, one for each of the following categories: liquidity, profitability, leverage, solvency, and activity. [5]

LDA provides an assessment of corporate credit quality using a discriminant function that classifies corporate borrowers into groups (default and non-default) based on their characteristics. Using sample observations, the set of parameters (β_i) is estimated for each independent variable (X_i): the discriminant score Z_i, obtained as the product of X_i and the β_i parameters, allows the mapping of firms into (non-continuous) default probability classes. This leads to the following equation:

$$Z_i = \sum_{i=1}^{n} \beta_i \cdot X_i \qquad (1)$$

LDA has been extensively use by practitioners, with a fairly good performance in predicting bankruptcy and other types of distress of non-financial firms worldwide (Altman, 1968; Deakin, 1972; Altman, 1983; Micha, 1984; Altman *et al.*, 2017). Over time LDA has also attracted criticism on some of its underlying assumptions, mainly based on the following arguments: failing firms and non-failing firms belong to two different populations; the choice of the normal distribution for the independent variables may not be appropriate; the covariance matrices for the two populations may differ. Besides, in LDA the standardized coefficients cannot be interpreted like the slopes of a regression equation and, consequently, the coefficients do not represent the relative weight of the variables. With these caveats in mind, researchers and practitioners have employed new methodologies to estimate the probability of default; in particular, Ohlson (1980) uses a conditional logit model to predict corporate default.

The logistic regression model (logit) provides estimates of the continuous probability of default from firms' observable characteristics using the two extreme values for the probability of default as the

[5] The original Z-score model uses the following ratios: working capital/total assets, retained earnings/total assets, EBIT/total assets, market value equity/book value of total debt, and total debt and sales/total assets.

dependent variable: 0 for financially sound firms and 1 for defaulted firms. The model assumes that firms belong to the same population and that a known structural relationship (additive and linear) exists between the observable characteristics of the firm and the probability of default, as in the following equation:

$$\ln\left(\frac{PD}{1\text{-}PD}\right) = \alpha + \sum_{i=1}^{n} \beta_i \cdot X_i + \varepsilon \qquad (2)$$

where PD is the probability that a default event occurs e.g. within the next 12 months, and ln (PD/(1 - PD)) is the log-odds ratio (or logit). The above expression can be solved for PD as follows:

$$PD = \frac{1}{1 + e^{(\alpha + \sum_{i=1}^{n} \beta_i \cdot X_i)}} \qquad (3)$$

The parameters β_i of the model are usually estimated with the maximum likelihood procedure to find the smallest possible deviation between the observed variable and the predicted variable.[6]

The logit has two practical advantages: first, it relaxes the restrictive assumptions of LDA, such as multivariate normality and equal covariance matrices; second, it can work with unbalanced samples. From a statistical point of view, the logit regression fits well the characteristics of the default prediction problem, where the dependent variable is binary (failure/non-failure) and it has the important advantage of ease of calibration of its results, by delivering a score between zero and one, which corresponds to the probability of default of the borrower. Another advantage is the interpretation of the outcomes: the estimated coefficients show the economic significance of each of the independent variables in the explanation of the PD.

These features have attracted a strong interest for the logit model from the academic literature (Gentry *et al.*, 1985; Mossman *et al.*, 1998, Altman *et al.*, 2013) and have prompted its adoption among practitioners. Logit is particularly fit for use in banks' IRB models. They should be intuitive (CRR, art. 179),[7] and the logistic regression guarantees an interpretable and quantifiable link between risk drivers and the default indicator. The logit model is also employed by several NCB in their ICAS models (Auria *et al.*, 2021).

In spite of its large diffusion, the logit model still has some limitations. They include: a) the lack of non-linear or complex interactions between observables and defaults; b) the sensitivity to outliers or missing data; and c) the difficulty to fully use large datasets owing to the intrinsic parsimony of the model.

[6] $\beta_{LOG} = \text{argmin}\left[\frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i \cdot \beta^T \cdot x_i + \ln\left(1 + e^{\beta^T \cdot x_i}\right)\right)\right]$

[7] For more information, see https://www.eba.europa.eu/regulation-and-policy/single-rulebook/interactive-single-rulebook/100971

4.2 Machine learning models

Machine learning involves teaching a computer routine to parse data, learn from them, and then make a prediction. The methodology 'learns' using large amounts of data and algorithms. The ML field has a long tradition, and recent improvements in data storage and computing power have made ML ubiquitous across many different fields and applications. As concerns credit scoring models, in recent years a growing strand of research indicates that the use of models based on ML algorithms for default prediction may provide a suitable alternative to statistical modelling. The ML approach to default prediction differs from the standard approach: 1) ML models are free from assumptions regarding their functional form, that is, they are non-parametric and allow for non-linear and non-monotonic modelling of the relationship between explanatory variables and outcomes; 2) ML algorithms are more computationally intensive, since they estimate multiple models and then select the best performing one using cross-validated accuracy measures. The discriminatory power of a model is typically assessed with the Area under the receiver operating characteristic (AUROC). AUROC is a measure of the ability of the model to assign higher PDs to firms that will default compared with financially sound firms. By construction, the AUROC statistic ranges from 0 ('the model is completely wrong') to 1 ('the model discriminates perfectly'), whereas 0.5 indicates a purely random model. In practice, a rating system with AUROC ≥ 0.7 is usually considered as adequate.[8]

ML algorithms can be grouped into two classes: supervised learning and unsupervised learning. In the first class, the algorithm learns from the associations between the category attributed to each outcome observation (for example default or non-default) and the observable characteristics for these observations. In turn the algorithm is able to predict the category an observation belongs to, given its characteristics (that is the algorithm performs a classification task). In the second class of algorithms, the observations are not grouped in different categories (i.e. an outcome variable is not provided) and the model puts forward a classification of the observations.

Several studies compare the performance of ML models and statistical models in corporate default prediction. Using a large dataset covering the North American corporate sector, Barboza *et al.* (2017) show that ML provides a significant improvement over statistical models. Bachman and Zhao (2017) compare the performance of ML models with Moody's proprietary regression model in the US corporate sector; ML models achieve an AUROC statistic which turns out to be 2-3 percentage points larger than that of the regression approach. Fantazzini and Figini (2009) compare the performance of a specific ML model (the random forest model, or RDF) with that of the logit model for the probability of survival using SME data in Germany. The study finds a weak association between in-sample and out-of-sample forecasting performance, revealing an overfitting problem that can be associated to ML. The logit model outperforms the RDF model in the out-of-sample forecast of default probability. Using a large dataset of Italian non-financial firms, Moscatelli *et al.* (2020) show that the use of ML techniques (RDF and gradient boosted trees, or GBT), can improve the performance of credit scoring models if the latter only rely on financial statement information. The value added of ML models declines when the scoring models also

[8] For illustrative purposes, suppose that the sample of firms displays an ex post frequency of default equal to 2 per cent. A model that assigns ex ante the largest (smallest) 2 per cent PDs to the same firms that eventually defaulted has an AUROC equal to 1 (0). See Fawcett (2004) and Chawla (2009).

rely on credit behaviour indicators based on the NCR. This suggests that ML models, which are relatively opaque, may be used as a benchmark for the probability of default obtained using more transparent statistical models.

The majority of researchers conclude that ML models, on average, outperform traditional statistical models thanks to their ability to capture more precisely indicators with a non-linear or non-monotonic relationship with default. This feature of ML models is very important for credit risk applications, but it comes at the cost of lower transparency relative to statistical models. ML models do not provide estimates of the parameters that relate predictors to the outcome variable, that is the models are non-parametric. Such 'black box' feature can make their rationale and forecasts difficult to grasp. For this reason, European banks have been hesitant to apply ML techniques to their IRB models for credit risk (EBA, 2021).

4.3 Hybrid models

A strand of the literature has explored hybrid credit scoring models that combine ML techniques with traditional statistical methods. Van Gestel *et al.* (2005) propose a model that combines the interpretability of the linear logistic regression with the flexibility and the ability to capture complex non-linearities of a support vector machine (SVM) model. Their results show that the hybrid model outperforms a simple statistical model. Dumitrescu *et al.* (2022) put forward a penalised logistic tree regression (PLTR) which tries to improve the predictive performance of the logistic regression through data pre-processing and short-depth decision trees. The hybrid model performs better than traditional statistical models while being comparable to machine learning methods.

5. The statistical model

5.1 Overview

In the development of the statistical model, we have tried to solve the trade-off between predictive power and transparency by combining features of traditional models and ML models. In particular, we preserve the simplicity and transparency of the logit model and, at the same time, we try to improve its predictive performance using ML techniques that allow for the inclusion of non-linear or non-monotonic relationships of the model variables with the default event.

A common type of non-linearity may arise from the existence of one or more univariate threshold effects on a single predictive variable (Dumitrescu *et al.*, 2022). ML decision tree methods (such as random forest) are powerful at detecting univariate threshold effects by using the non-linear relationship with default, but these methods lack interpretation. This opaqueness hinders credit scoring, especially when the final credit rating involves also expert analysis. Consequently, there is a trade-off between predictive performance and transparency. To solve this issue, one could seek an explanation for the decision-making

process of machine learning techniques (Cascarino *et al.*, 2022). Alternatively, one could preserve the simplicity of traditional models while improving their performance with the introduction in the model of variables with non-linear effects or univariate threshold effects (Van Gestel *et al.*, 2005, Dumitrescu *et al.*, 2022). Moscatelli *et al.* (2020) show that the value added of ML models declines when scoring models also rely on credit behaviour indicators, as in our case. To enhance the comprehension of the statistical model by credit analysts, we decided to follow the second approach, leaving the first one for future research. Therefore, we try to improve the discriminatory power of the logit model by employing, for certain variables, a discretized transformation obtained with the use of several decision trees (section 5.4.2).

The model draws from a pre-existing model, in operation from 2013 to 2022 (see Giovannelli *et al.*, 2020). For the estimation of the new model we started by checking whether each variable in the previous model still significantly contributes to default prediction. We also tried completely new variables. The development of the new version of the statistical model pursues the following objectives:

- improving the discriminatory power compared to the previous model;
- increasing the stability and consistency across different phases of the economic cycle;
- enhancing the transparency and including financial ratios that are widely recognized as explanatory of company credit-worthiness based on experience and the recent empirical literature.

The new model tries to achieve the above objectives with the introduction of some new features compared to the previous model. First, to increase stability and consistency, the observation period for the development of the model has been extended. The estimation sample considers six years of default history (from 2014 to 2019): four years are used as the training sample (from 2015 to 2018) and two years form the test sample (2014 and 2019).[9] By doing so, we use for both estimation and testing a suitable mix of 'good' and 'bad' years, which are representative of the likely range of variability of one-year default rates. The default data for 2020 are excluded from the estimation analysis owing to the extraordinary nature of that year, with the spreading of the Covid-19 pandemic and the ensuing government support measures. Nevertheless, we assess the model performance in terms of discriminatory power and backtesting also for 2020 (see section 7).

Second, separate models for different sectors usually improve predictive performance, as they allow for the selection of the most appropriate sector-specific financial ratios (Hajek, 2012; Lee and Choi, 2013). We find supporting evidence for this hypothesis, and apply it to the following macro-sectors that were also employed in the pre-existing model: industrials, trade, construction, services, real estate, and holdings. To further improve predictive performance, we apply a new distinction for the type of financial statement (ordinary financial statement or simplified financial statement).[10] Usually, simplified financial statements do not provide the breakdown of debt between financial and other liabilities and, consequently, the use of financial ratios including financial debt is precluded without making arbitrary assumptions. The distinction enables us to use some ratios that are widely employed in the literature and in expert analysis

[9] The previous version of the statistical model was developed using a two-years' time period (Giovannelli, *et al.* 2020).

[10] The simplified accounts may only be employed by joint stock companies that, for two consecutive financial years, have not exceeded any two of the following limits (micro and small enterprises): 1) total assets = EUR 4.4 million; 2) total revenues = EUR 8.8 million; 3) average number of employees during the year = 50 persons.

(such as the ratio between equity and net financial debt). As a result, the number of financial statement sub-models increases from six to eleven (for the holdings we developed only one model due to the small number of firms in this sector).

Third, the firm-bank relationship and credit availability differ significantly across firm size (Angori *et al.*, 2020; Kosekova *et al.*, 2023). Our analysis confirms the important role of size in the credit behaviour of non-financial firms. We thus estimate separate credit behaviour models according to company size.[11]

Fourth, we use a discretized transformation for certain variables in order to catch non-linear relationships with default (section 5.4.2) and we introduce new risk areas, such as business development in the financial statement sub-models and quality of credit receivables in the credit behaviour sub-models.

Finally, the updated Bank of Italy's statistical model consists of a system of logit models with two independent components providing separate credit scores:

- a financial component, employing a logit regression on yearly financial statement data, such as the debt sustainability ratio, financial structure, and liquidity ratios. This component consists of eleven sub-models, that account for different sectors and types of financial statement;
- a credit behaviour component, employing a logit regression on data from the NCR, which hinges on the credit record of each company. Three sub-models are set up for different firms considering their size (micro, small and medium-large) based on the European Commission definition (Recommendation 2003/361/EC).[12]

In line with banking industry practices, the two components are estimated separately and then merged into the final model through a further logistic regression that yields the final score. This is transformed into a PD via the inverse logit function:

$$PD = \frac{1}{1 + e^{-(\text{final score})}} \tag{4}$$

This approach is used to aggregate different information in the rating system. Under a hypothetical single model, the potentially high correlation between accounting variables and credit relationship variables could generate biased results, with a cost in terms of predictive power. Besides, the different timing of data (monthly data for NCR and annual data for the financial statement) and time lag of the data sources (up to 12 months for financial statements, 2 months for NCR) could generate distortion in the PD estimates (Giannozzi *et al.*, 2013).

The final stage of the model, namely the integration of the two components, is carried out with four models by size (micro, small, medium, and large firms). This distinction allows the relative importance of financial information and credit behaviour information to change with firm size.

[11] In the previous model, the size of a firm was approximated by the financial exposure towards the banking system. In a limited number of cases this classification could generate a distortion. In particular, if a large firm had a low financial exposure reported in the NCR it was classified as a small company.

[12] For more information, see https://ec.europa.eu/growth/smes/sme-definition_en

The structure of the model is presented in Figure 3.

Figure 3 – Statistical model architecture

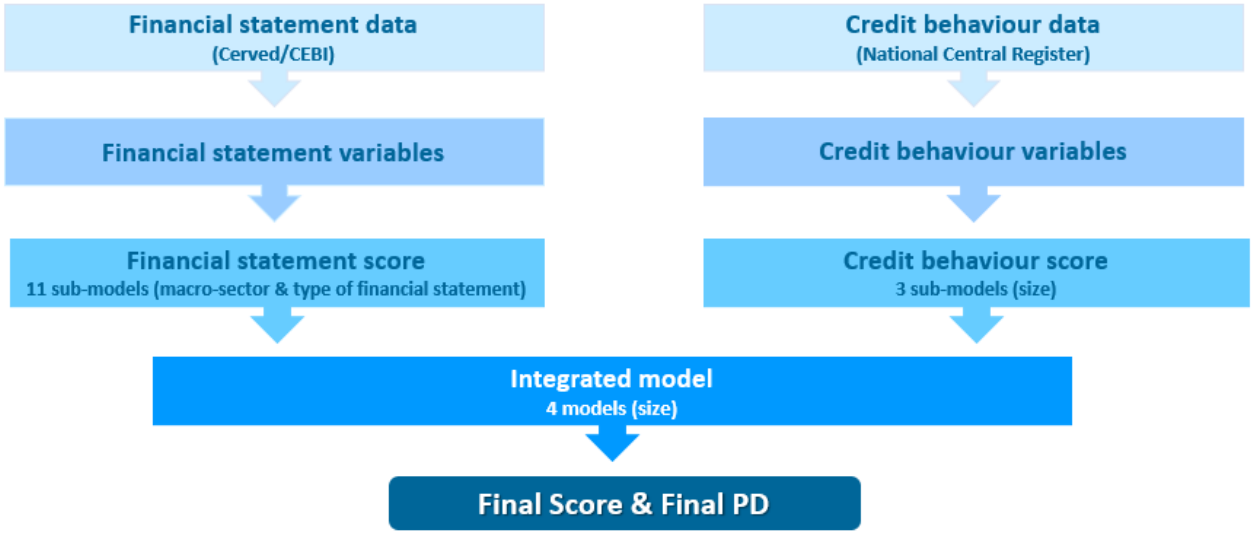
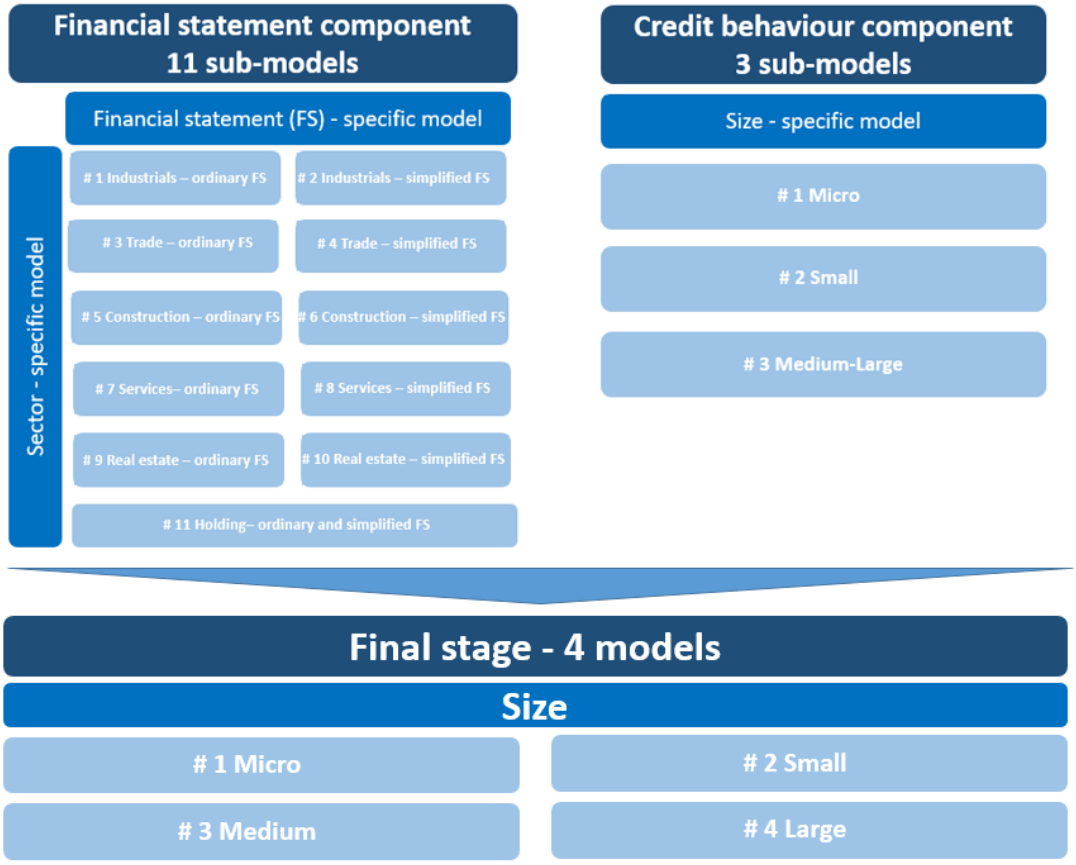


Figure 4 provides a summary of the sub-models. The estimation of separate sub-models, a common practice in the banking industry, enables us to select the most appropriate variables taking into account the characteristics of different firms. Sector-specific financial statement sub-models improve the discriminatory power (Hajek, 2012; Lee and Choi, 2013), while the breakdown according to the type of financial statement improves the usability of the model by the credit analysts.[13] For the credit behaviour component, the estimation of size specific models reflects the importance of company size in the credit relationship with the banking system.[14] For each sub-model we perform a variable selection process (described in Appendix 1).

[13] The expert assessment starts with an analysis of the consistency and validity of the data for the statistical model. Next, the assessment involves possible adjustments to the data (e.g. finance leasing is not included in financial debt according to Italian GAAP financial statements).

[14] In the credit behaviour component we consider medium and large firms together to increase the statistical significance of some ratios, owing to the low number of defaults among large firms. In the final stage of the model, though, we separate large and medium firms because the relative importance of financial information and credit behaviour information changes widely according to firm size (see section 5.5.3).

Figure 4 – Statistical model overview



5.2 Default definition

To identify the default status of the firms in our sample, in line with ECAF standards, we perform a merger of reporting information from the whole banking system. The information is provided by banks via AnaCredit[15] and the NCR, through which banks must report defaults according to Regulation (EU) No 575/2013 (CRR). The ICAS default definition relies on Article 178 of the CRR, which sets forth that a default occurs when a bank considers that the obligor is unlikely to pay (UTP) credit obligations or the obligor is past due more than 90 days on any material exposures to the bank. To obtain the default status, the whole default information for a given obligor is aggregated into a single indicator.

For the estimation of the model we use the binary default definition, whereby a borrower is considered in default if both of the following conditions are met:

- the total amount of exposure reported as bad debt, unlikely to pay, and more than 90 days past due by each bank is greater than 5 per cent of the total exposure of the borrower towards the banking system (materiality rule) and greater than EUR 500;
- the previous condition is met for three consecutive months (persistency rule).

[15] Analytical Credit (AnaCredit) is a dataset containing detailed information on individual bank loans in the euro area, harmonized across all Member States. The AnaCredit dataset project was launched in 2011 and data collection started in September 2018. On 18 May 2016 the ECB adopted Regulation (EU) 2016/867 on the collection of granular credit and credit risk data (known as 'AnaCredit Regulation'), following the principles approved by the Governing Council in 2015.

This definition of default is used also for calibration and internal validation of the model. For estimation purposes, we adopt the binary default definition for prudential reasons and owing to the choice of the logit model, that involves a binary independent variable.

As of January 2020, the new harmonized definition of 'fractional default' has been introduced for evaluating the performance of ICAS models in the yearly ECAF monitoring process. The new default definition aggregates the whole default information into a single default indicator. Two thresholds are applied:

- a materiality threshold of 2.5 per cent of the defaulted amount over the total exposure of the debtor;
- persistence for three consecutive months.

Over a given monitoring period, fractional default is equal to the maximum proportion in default which fulfils the two above conditions. In the case of bankruptcy, insolvency, judicial administration or similar measures (legal default) the materiality is considered equal to 100 per cent and no persistency is required. This definition is used for internal and external validation.

5.3 Data

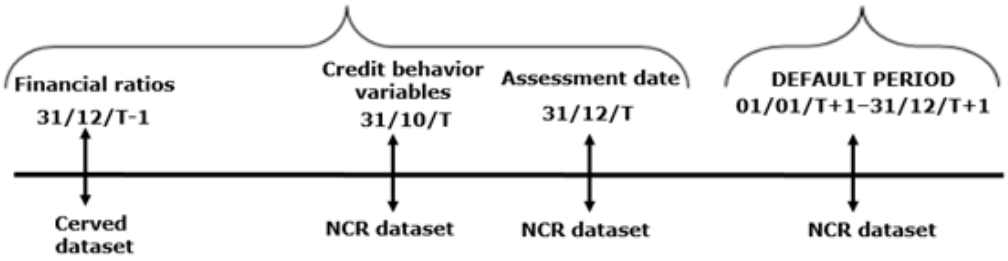
The model relies on two types of information: 1) financial statement data, and 2) credit behaviour data.

For financial statement data, we employ the Bank of Italy's financial statement archive which stores data collected by Cerved Group. Such data are available in two datasets: *Centrale dei Bilanci* (Cebi) and Cerved. The Cebi dataset contains around 80,000 financial statements per year covering nearly all the medium-sized and large firms and about half of the small ones. These data are collected partly through banks participating in the Cebi program and for the rest through the National Official Business Register (NOBR). Cerved manages a more comprehensive dataset of the Italian corporate sector that includes nearly all the small and micro limited-liability firms. Data are provided by NOBR. Financial statement data are available according to the Cebi reclassification accounting scheme, which can be applied to both the national GAAP and IFRS financial statements.

The NCR is a database with granular information on households' and firms' debt towards the banking and financial system. On a monthly basis, banks and other financial intermediaries are legally required to send to the central register a wide set of information about the financial liabilities and payment behaviour of individual firms and households. In return, banks have access to information on the debt exposure of their borrowers towards the whole banking system. The NCR collects information about all credit relationships with a minimum size of EUR 30,000. We use the NCR data for the default definition and the estimation of the credit behaviour model.

For data collection we use a cross-sectional approach: default data and independent variables refer to different time periods (Fig. 5). For each year T, we start by selecting limited-liability firms reported to the NCR at the end of the year and that are not in default. Then, we collect data according to the respective reference date, that is the most recent date prior to the reference date. The credit behaviour data have a two months' lag (31 October of year T), and the financial statement data are those for year T-1. The default monitoring period spans the following 12 months (from 1 January to 31 December of year T+1).

Figure 5 – Data timeline



For the development of the model, we used an extensive dataset of financial and credit behaviour indicators for Italian non-financial firms for the years 2014-19. Our dependent variable, binary default (section 5.2), reflects a system-wide definition of the non-performing status of a borrower. The historical default rate gives an aggregate measure of credit risk, which we model at firm level (Table 2). Credit risk peaked in 2014, in the aftermath of the European sovereign debt crisis and of the ensuing slowdown of the Italian economy. Following the monetary measures undertaken by the European Central Bank, the gradual improvement in the business cycle, and the exit of vulnerable firms from the market, aggregate default risk constantly declined, to below 3 per cent in 2019.

As stated earlier, the use of the binary default definition introduces a suitable degree of conservatism in the model estimation compared to the use of the fractional default definition (Table 2). In the remainder of this section we will only refer to the binary default definition.

Table 2 – Historical default rate

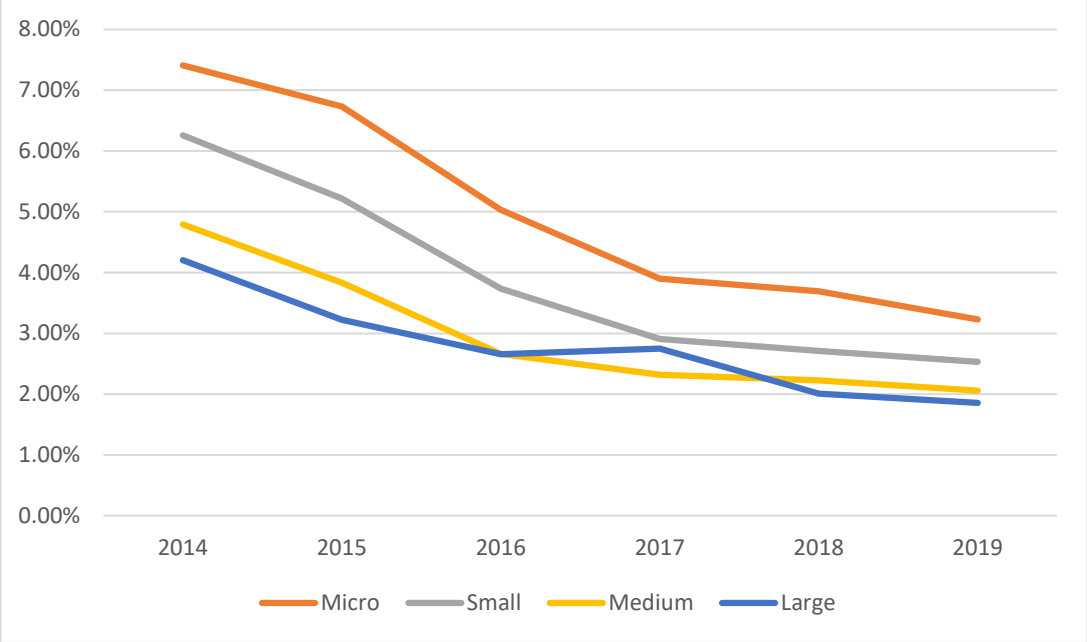
Period	Year	Number of firms	Binary default rate (%)	Fractional default rate (%)
Out-of-sample	2014	255,189	7.01	5.35
In-sample	2015	232,937	6.23	4.65
In-sample	2016	257,590	4.64	3.61
In-sample	2017	257,328	3.61	2.81
In-sample	2018	257,461	3.34	2.60
Out-of-sample	2019	267,341	2.96	2.36

Note: Our calculation on NCR data.

The estimation sample covers the years 2015-2018, while the test sample consists of the years 2014 and 2019. The latter are chosen to evaluate the performance of the model in a year with a higher default rate (2014) compared to the estimation sample and a year with a lower default rate (2019). The choice of four years as estimation sample and the large number of firms ensure a robust estimation, using a suitable mix of 'good' and 'bad' years.

Structural differences in default risk levels are usually associated with firm size and sector; consistently, credit risk models in the first place relate the firm PD to these factors (Altman *et al.*, 2017; Dwyer and Wang, 2011). Medium and large firms record lower default rates compared with micro and small firms and these differences become wider in periods of economic distress (Fig. 6). In 2014 and 2015 the differences in default risk by size class are at their peak, and then gradually decline with the improvement of the economic cycle.

Figure 6 – Default rate by size



Note: Our calculation on NCR and Cerved data. Binary default definition.
Firm size classes are defined according to the European Commission definition.

The lower riskiness of larger firms can be explained by their larger market shares and better profitability (measured e.g. by return on assets), leverage (measured by the ratio between equity and total debt), and debt sustainability (measured by interest expenses over EBITDA or sales; Table 3).

Table 3 – Key firm indicators by size

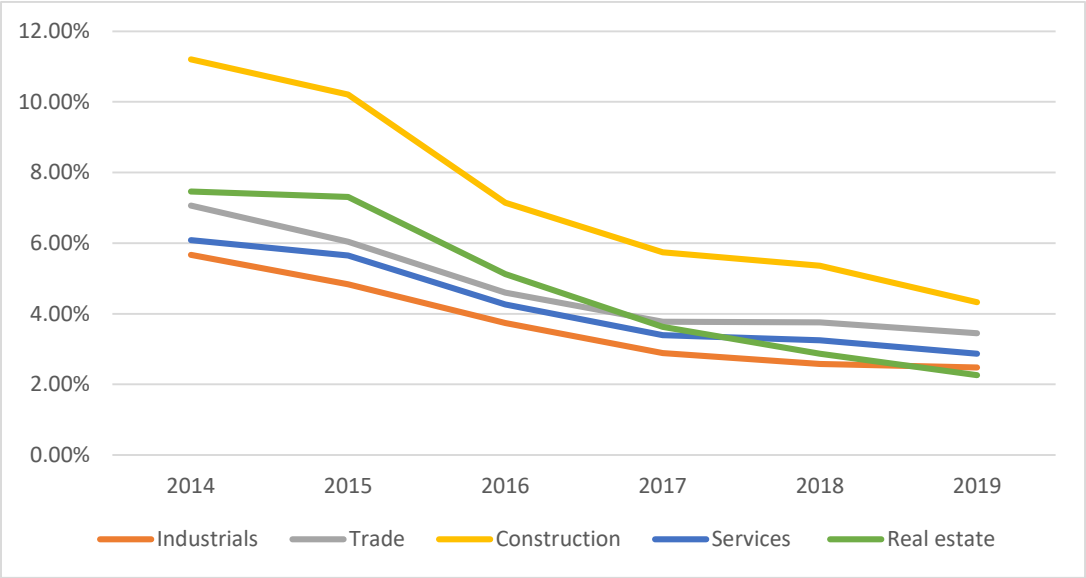
(2014-2019, median value)

Size	Number of firms	Total Assets (EUR mln)	Net Sales (EUR mln)	ROA (%)	Equity / Total debt (%)	Interest Expenses / EBITDA (%)	Interest Expenses / Sales (%)
Micro	1,100,481	0.8	0.5	3.2	18.8	19.3	2.3
Small	317,288	3.6	3.8	3.6	26.9	15.0	1.0
Medium	88,164	15.9	17.0	3.6	39.1	11.2	0.7
Large	21,913	87.8	89.5	3.5	44.4	9.4	0.7

Note: Our calculation on Cerved data.
Firm size classes are defined according to the European Commission definition.

The economic sector is another important dimension of firm credit risk, with more cyclical sectors (usually investment goods and construction) leading to more volatile revenues and greater fragility. Indeed, construction presents a significantly higher default rate than other sectors, such as industrials and services (Fig. 7). The latter sectors usually present more diversified businesses with greater exposure to foreign markets.

Figure 7 – Default rate by macro-sectors



Note: Our calculation on NCR and Cerved data. Binary default definition.
Firm economic sector affiliation is based on Cerved data.

Construction firms show on average a lower profitability and capitalization, and a worse debt servicing capacity (Table 4).

Table 4 – Key firm indicators by macro-sectors

(2014-2019, median value)

Sector	Number of firms	Total Assets (EUR mln)	Net Sales (EUR mln)	ROA (%)	Equity / Total debt (%)	Interest Expenses / EBITDA (%)	Interest Expenses / Sales (%)
Industrials	437,728	1.7	1.4	3.9	26.4	15.7	1.4
Trade	326,731	1.2	1.5	3.5	18.8	20.8	1.1
Construction	169,683	1.5	0.8	3.1	17.6	24.0	2.4
Services	397,260	0.8	0.7	3.7	19.7	14.7	1.6
Real Estate	183,765	1.6	0.1	1.7	30.7	20.0	11.2

Note: Our calculation on Cerved data.
Firm economic sector affiliation is based on Cerved data.

Within the same macro-sector, firms subject to different accounting standards (ordinary financial statement or simplified financial statement) show significant differences in terms of total assets and net sales (Table 5).

Table 5 – Firm distribution by macro-sectors and financial statement type

(2014-2019, median value)

Sector	Financial Statement type	Number of firms	Total Assets (EUR mln)	Net Sales (EUR mln)	ROA (%)	Equity / Total debt (%)	Interest Expenses / EBITDA (%)	Interest Expenses / Sales (%)
Industrials	ordinary	231,636	3.0	3.0	3.9	31.7	13.9	1.2
	simplified	206,092	1.0	0.9	3.8	21.2	17.9	1.8
Trade	ordinary	177,208	1.9	2.9	3.6	20.6	19.7	0.9
	simplified	149,523	0.8	0.9	3.4	16.7	22.4	1.4
Construction	ordinary	77,929	2.1	1.2	3.2	20.7	22.1	2.0
	simplified	91,754	1.1	0.6	3.0	15.3	25.9	2.9
Services	ordinary	153,300	1.4	1.4	3.7	21.9	13.2	1.2
	simplified	243,960	0.6	0.5	3.7	18.3	15.7	1.9
Real Estate	ordinary	41,971	2.6	0.2	1.4	31.4	19.3	11.7
	simplified	141,794	1.4	0.1	1.7	30.5	20.4	9.8

Note: Our calculation on Cerved data.
Firm economic sector affiliation is based on Cerved data.

5.4 Variables

As shown in section 5.1, the statistical model has a financial statement component and a credit behaviour component, each with a suitable sub-model distinction. The distinction in the first component allows for the selection of financial ratios based on the sectoral specific characteristics and of the information on financial liabilities, available only in the ordinary financial statement. The model distinction implicitly takes into account the firm size, as firms with an ordinary financial statement are larger than those with simplified reporting (Table 5).

The credit behaviour model distinction captures the structural differences in the use of credit lines and the relationship with the financial sector by firms of different size.

For each sub-model, the best discriminatory function is selected by means of the methodology described in Appendix 1.

The two components are merged via a logistic regression according to firm size.

The remainder of this subsection reviews the rationale for the selection of variables for each component and provides summary statistics.

5.4.1 Financial statement variables

The empirical literature employs a large variety of accounting ratios to predict firm default. Chen and Shimerda (1981) note that nearly 50 different financial ratios are significant in at least one empirical study.

They group the ratios in several risk areas and note that the inclusion of correlated variables as regressors distorts the relationship between variables and default.

Consistently with previous research, and in line with financial industry practice, we start by defining a long list of over 90 variables (considering all the 11 sub-models) grouped in the following risk areas:

profitability - profitability has a negative relationship with the PD (Altman, 1968; Altman *et al.*, 1977; Ohlson, 1980; Campbell *et al.*, 2008). We test over 20 variables including net profit and loss, EBITDA, EBIT, net sales, and cash flow at the numerator and total assets, tangible assets, and net sales at the denominator;

capitalization and financial structure - high leverage increases the riskiness of a company. (e.g. Bonaccorsi *et al.*, 2015). We try several variables including the ratios between equity or tangible equity, and total debt, net debt, total financial debt, or net financial debt (the latter variables apply only to the sub-models for firms with ordinary financial statements);

debt sustainability - poor debt sustainability reduces the credit worthiness of a company. In this risk category we try over 30 indicators including the ratios that measure the firm's ability to generate net sales, EBITDA, EBIT or cash flow to cover interest payments or some other measure at the denominator, such as total debt or financial debt. De Socio and Michelangeli (2017) measure firm vulnerability using the ratio between interest expenses and EBITDA. Other studies (Beaver, 1966; Blum, 1974) investigate the discriminant power of the ratio between EBITDA or cash flow and financial debt, showing a positive relationship with credit soundness. In addition, for ordinary financial statements, we include also interest rate risk exposure such as return on debt (ROD), measured as the ratio between interest expenses and average financial debt. Based on Modigliani and Miller (1958), we expect a direct relationship between ROD and PD;

liquidity - high liquidity reduces the PD. Altman and Sabato (2007) find that the PD in the US economy is negatively related to the ratio of cash over total assets. This risk area measures the extent to which a company has liquid assets relative to the size of its current or total liabilities. We consider over 10 variables including cash and marketable assets over total debt or financial debt, and traditional variables such as the current ratio (Beaver, 1966; Altman *et al.*, 1977), quick ratio (Giannozzi *et al.*, 2013), and other ratios between assets and liabilities such as financial mismatch;

activity and efficiency - the firm's operating efficiency has an impact on the PD. Following several studies (Alberici, 1975; Luerti, 1992; Piatti *et al.*, 2015), we test the discriminatory power of several cash conversion cycle variables, such as days receivables outstanding, days inventory outstanding and days payables outstanding;

business development - the variables in this group capture the stability and trend of firm performance. We consider sales growth, total asset growth, change in value added and equity;

size and age - large firms default less often (Fig. 6). Charalambakis and Garret (2019) find that size is negatively correlated with the probability of default of Greek private firms. Other authors find that the logarithm of age is negatively correlated with the PD (Altman *et al.*, 2010; Antunes *et al.*, 2016). We

measure size using the natural logarithm of total assets or net sales; we measure age by its natural logarithm.

Following the procedure described in Appendix 1, for each financial statement sub-model we define the final combination of variables and the relative weights (Tables 8 and 9). Table 6 provides summary statistics for default and non-default firms for the variables selected for at least one of the eleven financial statement sub-models.

Table 6 – Financial statement variables

(2014-2019)

Risk area	Variable	Mean (non-default)	Mean (default)	Median (non-default)	Median (default)
Profitability	Cash flow / Net sales (%)	8.1	1.3	5.1	2.5
	Cash flow / Total assets (%)	5.2	1.8	4.3	1.6
	Value Added / Total assets (%)	25.9	17.5	20.3	12.1
Capitalization & financial structure	Equity / Financial net debt (%)	245.7	98.6	79.3	29.4
	Equity / Total net debt (%)	61.9	27.4	24.6	9.9
Debt sustainability	Interest expenses / EBITDA (%)	36.5	63.4	20.3	56.6
	Interest expenses / Cash flow (%)	84.4	166.6	31.3	136.4
	Interest expenses / Net sales (%)	5.5	10.1	1.6	3.7
	ROD (%)	5.8	7.1	4.0	5.4
	Net sales / Total net debt	1.9	1.1	1.4	0.8
Liquidity	Current ratio = (current assets / current liabilities) (%)	1.6	1.5	1.2	1.1
	Financial mismatch = [(current liabilities – current assets) / total assets] (%)	-0.1	-0.1	-0.1	-0.0
	Cash / Total short term debt (%)	17.2	6.9	4.9	1.7
Activity and efficiency	Days receivables & inventory outstanding	339	612	82	153
	Days receivables outstanding	84	112	54	63
	Days payables outstanding	520	654	140	194
Business development	Sales growth (%)	5.5	1.9	0.0	0.0
Age	Log(age)	2. 6	2.4	2.6	2.4

Note: our calculation on Cerved data.
The mean and median are computed after winsorization and transformation (see Appendix 1).

As expected, profitability measures, such as cash flow over sales, cash flow over total assets and value added over total assets, are associated with credit worthiness; low or negative values for these ratios can be a signal of future financial problems.

Firm capitalization has a negative relationship with default rate and probability of default. For firms with ordinary financial statements, we employ the ratio between equity and net financial debt (financial debt –

cash), commonly used by analysts to assess the level of capitalization of a company; for firms with the simplified financial statement, we employ the ratio between equity and net total debt (total debt – cash).

As concerns debt sustainability, interest expenses over EBITDA (or cash flow) and interest expenses over net sales show a positive relationship with default and a high discriminatory power. The numerator of these variables captures the effect of credit conditions, as it includes the impact of the level of debt and its cost, while the denominator reflects operating profitability and business activity, respectively. Within the same risk area, we also consider: the debt coverage ratio, measured by net sales over net debt, which shows a negative relationship with default; and ROD, which measures the cost of debt, and is available only for firms with an ordinary financial statement.

The ratio of cash over total short term debt and that of current assets over current liabilities increase the possibility to meet the firm's financial obligations on time; empirically, both variables show a negative relationship with default.

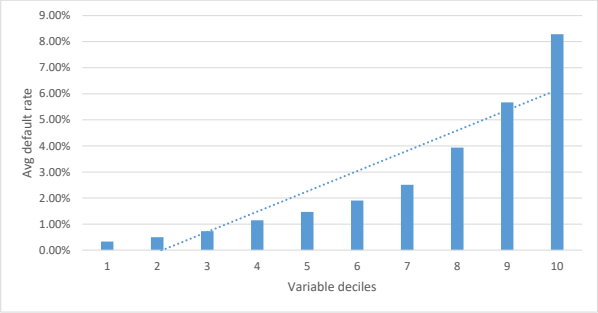
The age of firms displays a negative impact on PD, indicating a greater soundness of firms which stand in the market for longer.

The above variables generally display a linear relationship with the default rate. An example for industrial firms with the ordinary financial statement is shown in figure 8. The linear relationship of the above variables with default allows for their direct use as input to logit.

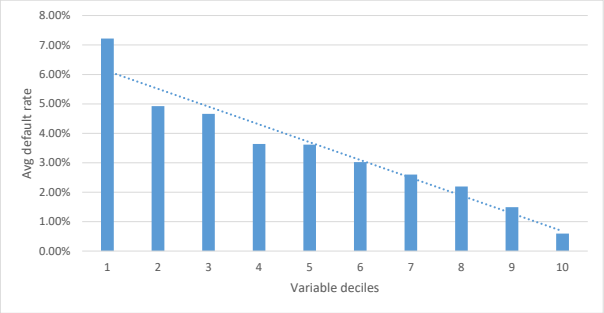
Figure 8. Default rate by deciles – linear relationship variables

(2014-2019, industrials, ordinary financial statement)

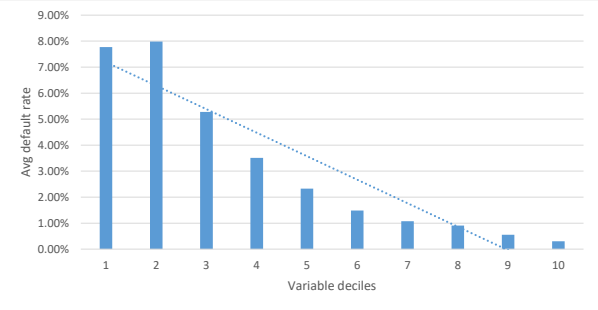
Panel 1. Interest expenses / EBITDA



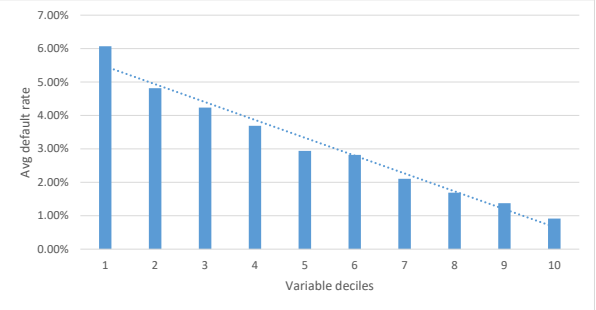
Panel 2. Equity / net financial debt



Panel 3. Net sales / total net debt



Panel 4. Current ratio



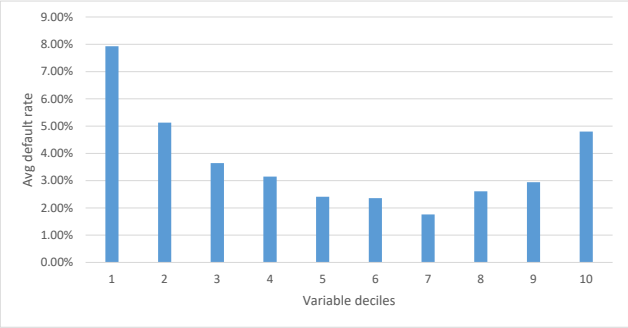
5.4.2 Nonlinearity

Some variables of interest from the financial statement exhibit a non-linear relationship with default risk in our sample. This is the case of e.g. sales growth (Fig. 9, panel 1). Other variables belonging to the 'activity and efficiency' risk area show a relationship with default rate that is non-monotonic for a large portion of the distribution. Nevertheless, we note that riskiness increases with a linear trend after reaching a certain threshold (Fig. 9, panels 2-4).

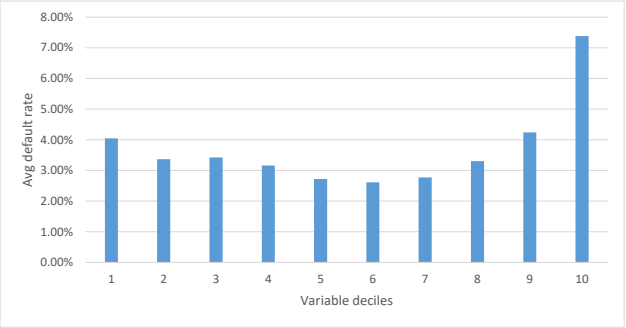
Figure 9. Default rate by deciles – non-linear relationships

(2014-2019, services, ordinary financial statement)

Panel 1. Sales growth

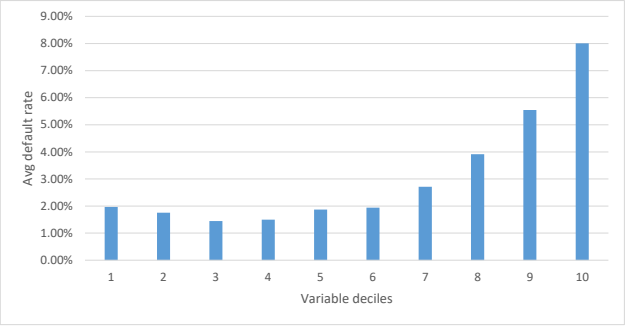


Panel 2. Days receivables outstanding

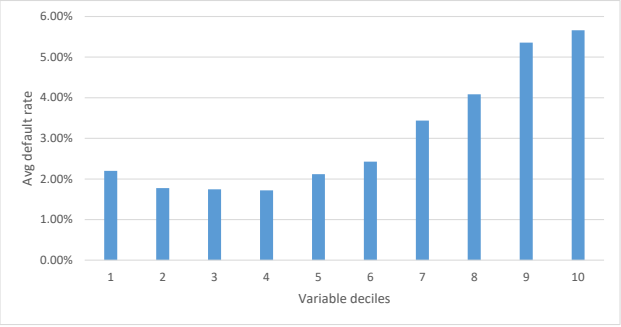


(2014-2019, industrials, ordinary financial statement)

Panel 3. Days receivables & inventory outstanding



Panel 4. Days payables outstanding



These non-linear variables cannot be accommodated in the standard logistic regression of default. Nevertheless, to fully use the financial statement information, we apply decision tree techniques to these variables and create new discretized variables as candidate indicators for the logit. Decision trees are partition algorithms that recursively split the dataset into smaller subgroups (or branches) that best separate defaulters from non-defaulters. At each iteration, the decision tree algorithm chooses a value for the classification variable, so as to minimize a measure of heterogeneity (impurity index) in the resulting branches with respect to the classification variable.[16] The process continues within each branch until a

[16] We use the Gini coefficient as a measure of impurity. Given a set of observations O with a binary value for each observation $y_i \in \{0,1\}$, the Gini impurity coefficient is defined as $GI(O) = 2 \cdot \sum_i p_i(1-p_i)$, where p_i is the percentage of observations in O such that $y_i=1$. The Gini impurity ranges between 0 ($p_i = 0$ or $p_i = 1$) and 0.5 ($p_i = 0.5$).

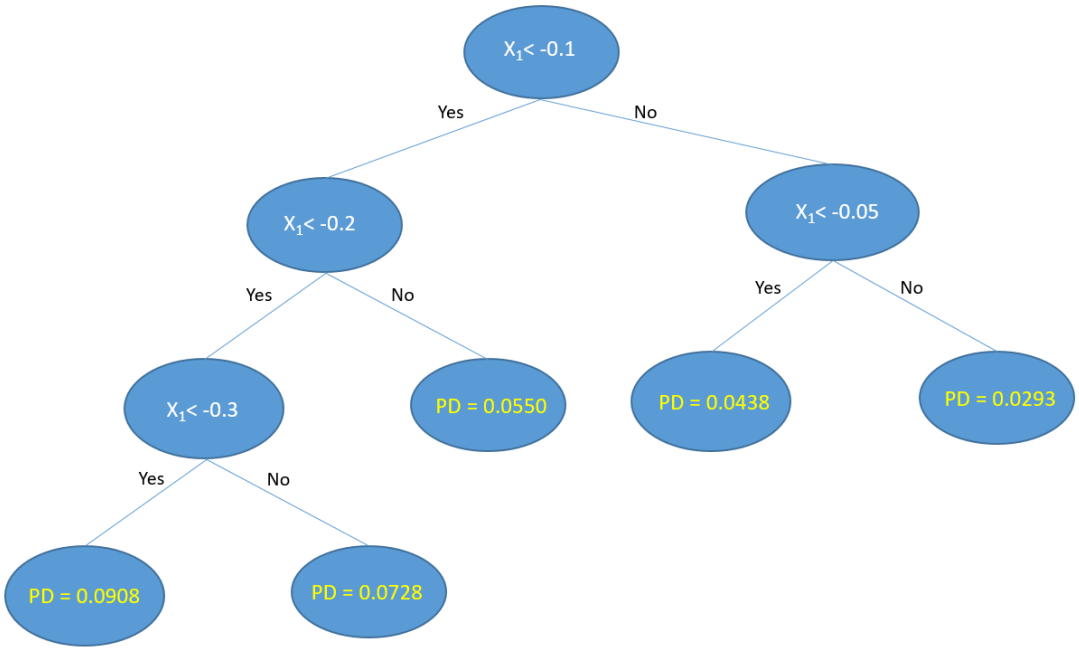
stopping condition is reached, such as a small number of observations in the branch or no significant reduction in impurity.

For illustrative purposes, we provide below a short description of the treatment on the four variables in figure 10.

As concerns sales growth, empirically both a large growth and a large decline lead to an increase in default rate. From a judgemental point of view, it is easier for a credit analyst to understand a negative effect on PD of a decline in sales rather than of their growth; hence we apply the decision tree to identify thresholds related only to negative sales growth.[17] As shown in figure 10, we apply a simple classification tree containing only the sales growth variable as the splitting parameter (X_1). The first level of the tree splits the whole sample into two middle branches depending on the value of sales growth. Firms with sales growth higher than -0.1 are further divided according to their sales growth value. Firms with negative sales growth up to -0.05 end up in a leaf with a PD equal to 0.0293, while firms with negative sales growth between -0.05 and -0.1 have a PD equal to 0.0438. On the other branch, firms with sales growth worse than -0.1 are divided between firms with negative growth up to -0.2 that end up in a leaf with a PD equal to 0.055, while firms with sales growth lower than -0.2 undergo a further split distinguishing those with sales growth lower than -0.3 (PD equal to 0.0903) and those with sales growth between -0.2 and -0.3, that end up in a leaf with a PD equal to 0.0728.

Figure 10. Sales growth, decision tree techniques

(2014-2019, services, ordinary financial statement)



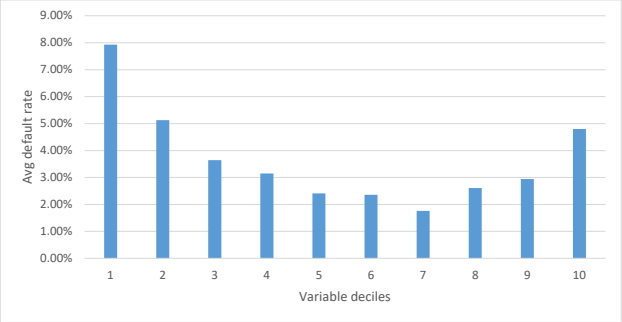
[17] All firms with positive net sales growth were considered as firms with a constant level of turnover (the variable sales growth was set equal to zero).

At the end of the procedure, the new discretized variable has a linear relationship with default, thus fitting the logistic regression model (Fig. 11).

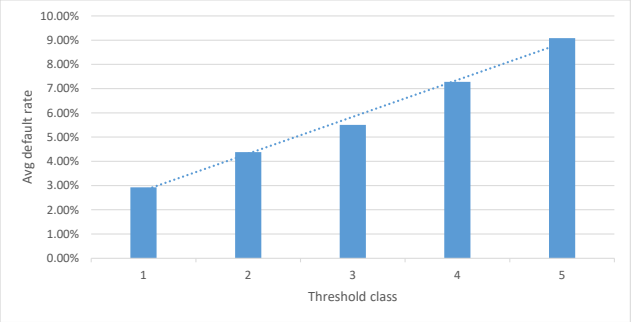
Figure 11. Sales growth

(2014-2019, services, ordinary financial statement)

Panel 1. Default rate by decile



Panel 2. Default rate by discretized class

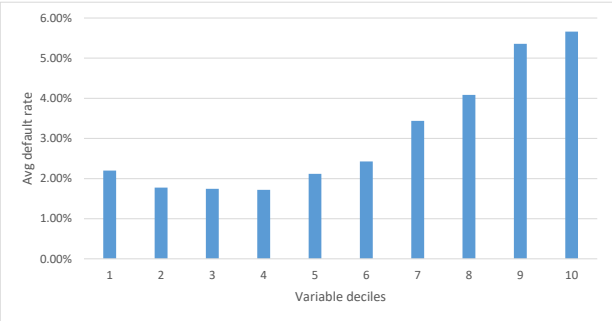


On the liability side, days payables outstanding (DPO) measure how long it takes on average for a company to pay its suppliers. A large DPO may be preferable for working capital management, as it may suggest favourable credit terms with suppliers. Nevertheless, after breaching a certain threshold, a large value of DPO is an indication of the company struggling to meet its obligations on time, thus revealing an increase in default risk. Consequently, the decision tree suggests to consider only the extreme part of the distribution. The resulting four discretized classes show a linear and positive relationship with default rate (Fig. 12).

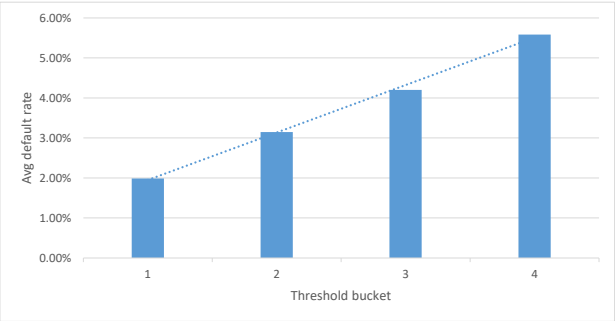
Figure 12. Days payables outstanding

(2014-2019, industrials, ordinary financial statement)

Panel 1. Default rate by decile



Panel 2. Default rate by discretized class



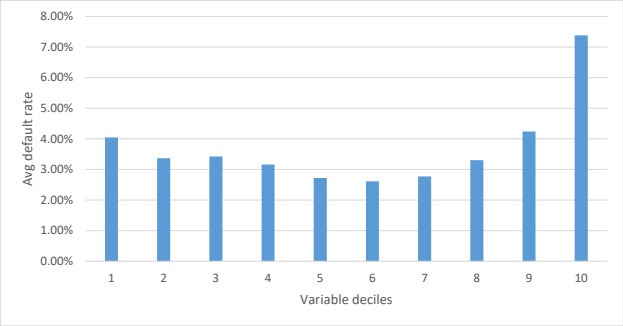
On the asset side, days receivables outstanding (DRO) is a measure of the average number of days for a company to collect payment from customers. A large number of days indicates an inadequate operating efficiency and/or customers who are not financially creditworthy. Late payment or the default of customers can have negative effects, such as cash flow shortages and economic losses. Days inventory outstanding (DIO) is the average number of days for the inventory to be sold. A high level of inventories could signal difficulties for a company in selling its products, with the risk of writing off a portion of the inventories. We use two discretized variables created from DRO and the sum between DRO and DIO, respectively. For the first ratio, the procedure identifies three classes with a positive relationship with default. For the second ratio, the discretized risk classes are four (Fig. 13).

Figure 13. Activity & efficiency ratios

Days receivables outstanding

(2014-2019, services, ordinary financial statement)

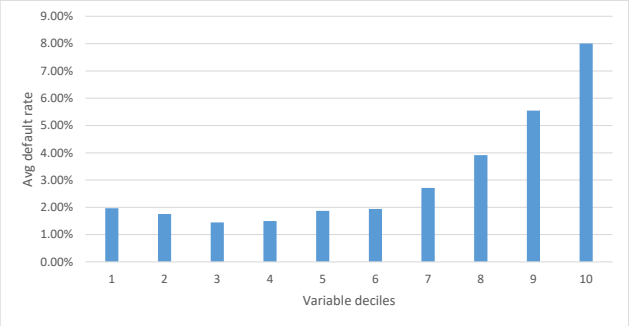
Panel 1. Default rate by decile



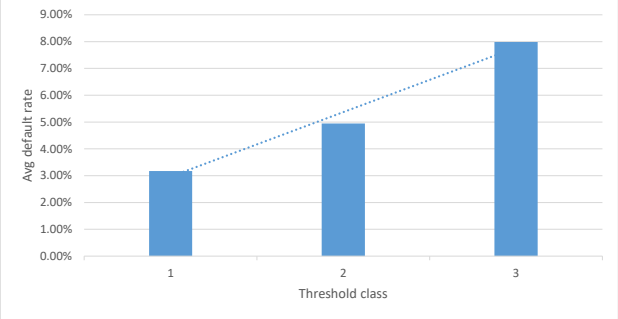
Days receivables & inventories outstanding

(2014-2019, industrials, ordinary financial statement)

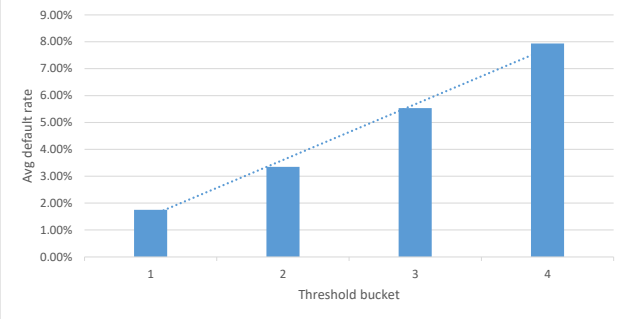
Panel 2. Default rate by decile



Panel 3. Default rate by discretized class



Panel 4. Default rate by discretized class



5.4.3 Credit behaviour variables

The NCR dataset is extremely granular, as it contains information about different credit lines held by each firm, distinguishing three classes of loans: term loans, current account revolving credit lines, and account receivables revolving credit lines. For each credit line, the dataset records drawn and undrawn amount, excess and unauthorized overdrafts, and default classification.

Several studies examine the role of credit behaviour in predicting default. Firms that draw more on their credit lines have a higher default rate (Jiménez *et al.*, 2009). Such firms may have liquidity issues, and high leverage or bad financial performance may make it difficult for them to acquire additional funding other than drawing down their credit lines (Zhao *et al.*, 2014). Credit line usage and excess overdraft are significant explanatory variables of default, especially for small business firms (Norden and Weber, 2010). More specifically, the utilization rate for current account revolving credit lines and excess overdrafts are good predictors of default for Italian SMEs and, more generally, the inclusion of credit behaviour variables improves the accuracy of credit scoring models (Giannozzi *et al.*, 2013).

Based on the existing empirical studies and the industry best practices, we define a list of over 70 variables for all the 3 sub-models, that cover the following risk areas:

utilization rate – this variable is defined as the ratio between drawn amount and granted amount, for each type of credit line and at various frequencies (last quarter, last six months, etc.);

financial distress – this is defined alternatively as the number of months, or absolute and relative amounts of overdrafts, for each type of credit line at various frequencies, or as past delinquencies and past default status. Previous studies show that excess overdrafts or past delinquencies are positively correlated with default (Norden and Weber, 2010, Gallucci *et al.*, 2022);

debt composition – this is measured by variables that describe the debt characteristics of a given company, including a dummy for the presence of medium-long term loans and the breakdown of total debt by different types of loans. Earlier research shows that longer maturities are associated with better financial conditions and lower informational opacity of the borrower, and that creditors use shorter maturities to enforce monitoring of riskier borrowers (Ortiz-Molina and Penas, 2008);

quality of credit receivables – the indicators in this area capture the quality of the credit portfolio used in account receivables revolving credit lines, like the ratio of unpaid credit receivables over the total amount of credit expires or over net sales. We expect that a large amount of unpaid credit receivables is associated with a large PD;

trend ratio – it is measured alternatively as the change of utilization rate, evolution of credit lines, change in the number of reporting banks, and number of first information requests, all of them at different frequencies. Previous evidence is that the overall effect of more concentrated banking relationships involves a higher probability of incurring into financial distress, due to an increase in liquidity risk (Carmignani and Omiccioli, 2007). In the last year before default, the number of reporting banks decreases (Salvadè *et al.*, 2021).

Table 7 provides summary statistics on a selection of the above variables with a breakdown for default and non-default firms.

Table 7 – Credit behaviour variables

(2014-2019)

Risk area	Variable	Mean (non-default)	Mean (default)	Median (non-default)	Median (default)
Utilization rate	Drawn amount/granted amount - current account revolving credit lines - average last three months (%)	40.8	78.6	33.0	95.9
	Drawn amount/granted amount – short term credit lines - average last six months (%)	27.7	49.7	1.24	60.1
	Drawn amount/granted amount – account receivables revolving credit lines - average last six months (%)	48.1	66.9	49.1	75.1
Financial Distress	No. of months (last six months) with overdraft on current account revolving credit lines	0.4	2.1	0.0	2.0
	No. of months (last six months) with overdraft on term loans	0.2	1.6	0.0	0.0
	Overdraft % on term loans, six months' average (%)	0.2	1.8	0.0	0.0
	Dummy default status (last six months)	0.0	0.1	0.0	0.0
Debt composition	Dummy no medium-long term credit lines	0.0	0.1	0.0	0.0
Quality of credit receivables	Unpaid/Expired amount on account receivables credit lines - average last six months (%)	12.3	31.0	1.2	14.4
	Unpaid amount on account receivables credit lines - average last six months / net sales (%)	2.3	10.2	0.0	0.4
Trend	Dummy reduction in the number of reporting banks – last six months	0.1	0.1	0.0	0.0
	No. of first information requests - last six months	0.4	0.5	0.0	0.0

Note: our calculation on NCR data.

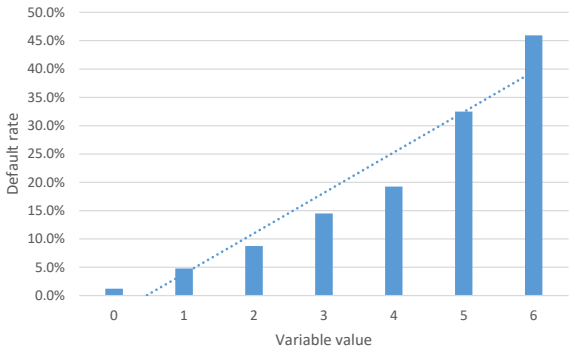
We find that the utilization rate variables, under various definitions, are significantly associated with default. As expected, financial distress variables are also clearly associated with future default.

The ratio of overdraft percentage on term loans has been discretized due to the skewness of its distribution, with only a small number of firms having an overdraft. The number of months with an overdraft on revolving credit lines and term loans, which is discrete and bounded by definition, exhibits a strong linear and monotonic relationship with default risk (Fig. 14).

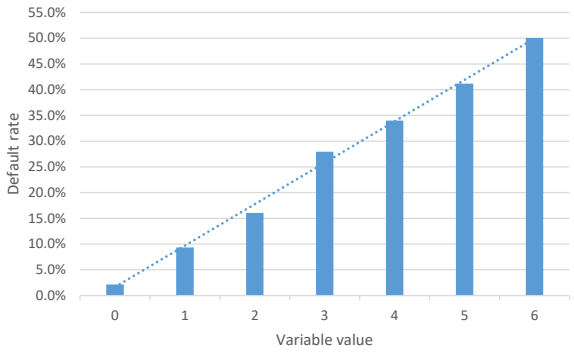
Figure 14. Overdraft analysis

(2014-2019, small firms)

Panel 1. No. of months w/overdrafts, account revolving credit lines



Panel 2. No. of months w/overdrafts, term loans



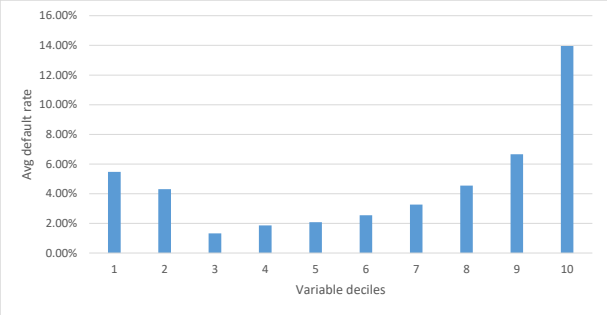
Consistently with Salvadè *et al.* (2021), we find that a decline in the number of banks in a semester is associated with a higher probability of default in the following twelve months. Furthermore, for micro and small firms, a large number of first information requests in the last six months is associated with higher riskiness. Similar to Ortiz-Molina and Penas (2008), our data show that the absence of medium-long term loans for micro and small firms is associated with an increase in PD.

As expected, we find that a bad quality of credit receivables has a positive relationship with default. Unrecoverable receivables have a direct impact on profit and liquidity. Within this area, we try measures of the general level of customer credit worthiness. We consider both a relative indicator, measured by the ratio between unpaid and expired amount on account receivables credit lines in the last six months, and an absolute indicator, namely the ratio between the average amount over the last six months of unpaid amount on account receivables credit lines over net sales. For both variables, we find a non-linear and non-monotonic relationship with default. Nevertheless, both variables exhibit an increase in default rate (with a linear trend) in the final part of the distribution, when a significant portion of the customer credit portfolios is not financially sound or when the amount of trade credit in default is above a certain amount compared to the size of the company, measured by net sales (Fig. 15, panels 1-2). Hence, with the decision tree technique we obtain two discretized variables with, respectively, five and three risk classes associated linearly with default (Fig. 15, panels 3-4).

Figure 15. Quality of credit receivables

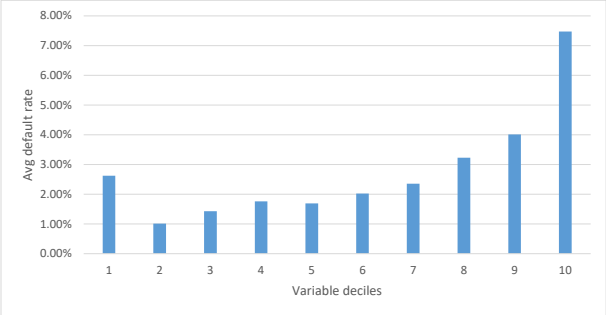
(2014-2019, micro firms)

Panel 1. Unpaid/Expired amount on account receivables credit lines avg six months

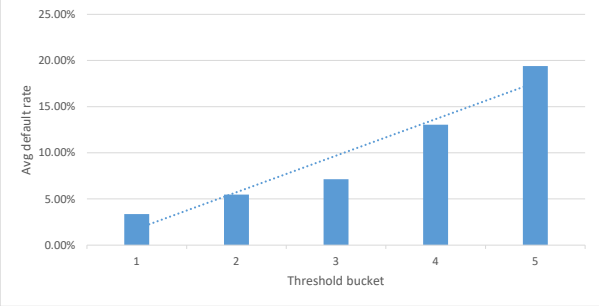


(2014-2019, medium-large firms)

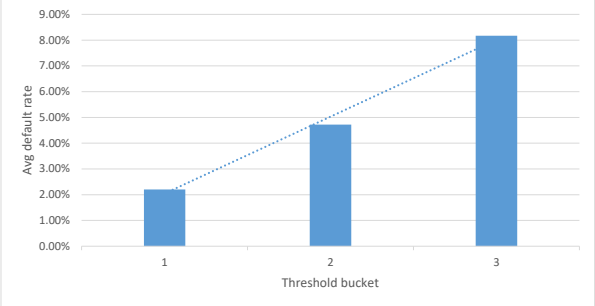
Panel 2. Unpaid amount on account receivables credit lines avg six months / net sales



Panel 3. Discretized variable



Panel 4. Discretized variable



5.5 Model development

5.5.1 Financial statement model

As shown above, the financial statement model consists of eleven sub-models based on the company macro-sector and type of financial statement. The breakdown according to the sector enables us to take into account the characteristics of different kinds of firms, while the type of financial statement enables us to fully use financial information (see section 5.1). For each sub-model we perform an independent variable selection process (see Appendix 1). We provide below a parsimonious representation of the main economic relationships estimated with each of the sub-models. These are summarised in Table 8 for firms with the ordinary financial statement and in Table 9 for those with the simplified financial statement. A '+' ('-') indicates a positive (negative) relationship of the variable with the default rate, i.e. whether an increase in the variable is associated with a higher (lower) PD. Variables with no sign are not considered in the specific sub-model.

Table 8 – Ordinary annual financial statement models

Risk area	Variable	Industrial sector				
		Industrials	**Trade**	**Construction**	**Services**	**Real Estate**
Profitability	Cash flow / Net sales	-	-		-	
	Cash flow / Total assets			-		
	Value added / Total assets					-
Leverage & financial structure	Equity / Net financial debt	-	-	-	-	-
	Fixed assets / Total assets					-
Debt sustainability	Interest expenses / EBITDA	+	+	+	+	
	Interest expenses / Cash flow					+
	Interest expenses / Net sales	+	+		+	
	Net sales / Total net debt	-	-		-	
	ROD (return on debt) = Interest expenses / Average financial debt	+	+	+	+	+
Liquidity	Current ratio = current assets / current liabilities	-		-		
	Financial mismatch = (current liabilities – current assets) / total assets		+		+	
	Cash / Total short term debt	-	-	-	-	-
Activity	Days receivables outstanding + Days inventory outstanding (discretized)	+				
	Days receivables outstanding (discretized)			+	+	+
	Days payables outstanding (discretized)	+	+	+		
Business development	Net sales negative variation (discretized)	+	+		+	
Age	Log(age)	-	-	-	-	

Table 9 – Simplified financial statement models

Risk area	Variable	Industrial sector					
		Industrials	Trade	Construction	Services	Real Estate	Holdings
Profitability	Cash flow / Net sales	-	-		-		
	Cash flow / Total assets			-			
	Value added / Total assets					-	
	EBIT / Total assets						-
	Delta returns						-
Leverage & Financial structure	Equity / Total net debt	-	-	-	-	-	
	Equity / Total assets						-
	Fixed assets / Total assets					-	
	Financial assets / Total assets						-
Debt sustainability	Interest expenses / EBITDA	+	+	+	+		
	Interest expenses / Cash flow					+	
	Interest expenses / Net sales		+		+	+	
	Net sales / Total net debt	-	-		-		
Liquidity	Financial mismatch = (current liabilities – current assets) / total assets	+	+	+	+		
	Cash / Total short term debt	-	-	-	-	-	
	Cash / Total debt						-
Activity	Days receivables outstanding + Days inventory outstanding (discretized)	+	+		+		
Business development	Net sales negative change (discretized)	+	+		+		
Age	Log(age)	-	-	-	-		

To estimate the relative weight of the variables, we use the standardized coefficient for each variable within a specific sub-model. We find that, on average, debt sustainability is the most important risk dimension, with the ratio interest expenses over EBITDA among the most significant variables for industrials, trade and services, for both ordinary and simplified financial statement sub-models. Besides, our results indicate that interest expenses over net sales is highly significant for trade firms, while net sales over total net debt is more important for services.

Leverage and liquidity are also significant risk dimensions. Within the liquidity area we find that the ratio between cash and total short term debt is a good predictor for each economic sector, and it is more important than other ratios that compare current assets with short term liabilities. In the leverage risk area, capitalization ratios are good predictors for capital intensive sectors, such as real estate and construction. We use the ratio between equity and net financial debt for the ordinary financial statement sub-models and the ratio between equity and total net debt for the simplified financial statement sub-models.

Overall, the distinction for the type of financial statement and the introduction of efficiency and business development variables with decision tree techniques significantly improve the discriminatory power of the sub-models in comparison with their previous version.

5.5.2 Credit behaviour model

The credit behaviour model consists of three sub-models (micro, small, and medium-large firms). The main relationships for the credit behaviour sub-models are summarised in Table 10. A '+' ('-') indicates a positive (negative) relationship of the variable with the default rate, i.e. whether an increase in the variable is associated with a higher (lower) PD.

Table 10 – Credit behaviour models

Risk area	Variable	Credit size		
		Micro	**Small**	**Medium-Large**
Average utilization rate	Drawn amount/granted amount - current account revolving credit lines - average last three months	+	+	+
	Dummy no-current account revolving credit lines - last three months	+	+	+
	Drawn amount/granted amount – short term credit lines - average last six months			+
	Dummy no-short term credit lines - last six months			+
	Drawn amount/granted amount – account receivables revolving credit lines - average last six months	+	+	
	Dummy no-account receivables revolving credit lines - last six months	+	+	
Debt composition	Dummy no medium/long term credit – last six months	+	+	
Financial distress	No. of months (out of last six) with overdraft on current account revolving credit lines	+	+	+
	No. of months (out of last six) with overdraft on term loans	+	+	+
	Overdraft % on term loans, six month average (discretized)	+	+	+
	Dummy default status – last six months	+	+	+
Quality of credit receivables	Unpaid/Expired amount on account receivables credit lines - average last six months (discretized)	+	+	
	Unpaid amount on account receivables credit lines - average last six months / net sales (discretized)			+
Trend	Dummy reduction in the number of reporting banks – last six months	+	+	+
	No. of first information requests - last six months	+	+	
Size	Net sales (discretized)			-

Consistently with an earlier study (Giannozzi *et al.*, 2013), according to the standardized coefficient methodology, the most important risk area for all the sub-models is the utilization rate and, in particular, the rate on the current account revolving credit lines. Unfortunately, these ratios are not always available due to some missing values, arising from the absence of short term loans or of other loan types. When this happens, we set the ratio equal to 0 (best possible value) and introduce a specific dummy variable which adjusts the credit score for the average riskiness of the sub-set of firms within the estimation sample for which this type of loans is not reported in the NCR.

The second most important risk dimension is financial distress, in particular the ratio constructed with the amount of overdraft on (discretized) term loans and the ratio that employs the number of months with excess overdraft on account revolving credit lines and term loans. The breakdown of the type of loan for overdraft indicators is one of the main improvements in comparison with the previous version of the credit behaviour model.

Other risk areas, such as the quality of credit receivables and the trend variables, are statistically significant but, as expected, their relative weights are less important compared to those of the average utilization rate and overdraft indicators. Nevertheless, the introduction of the former is the second most significant change compared with the previous model.

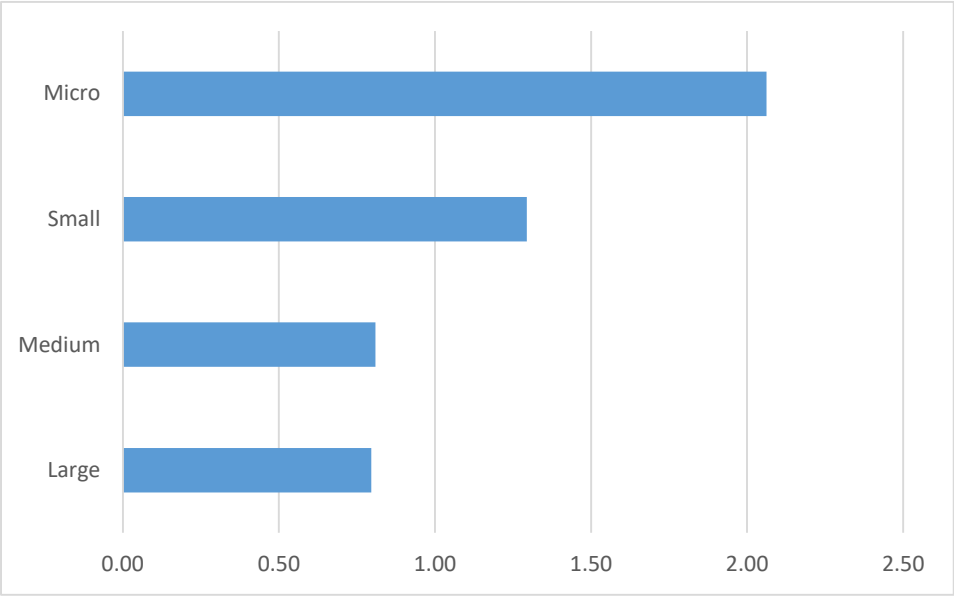
Within the medium-large firms' sub-model, we hypothesize that larger firms have a greater bargaining power towards their bank counterparties. The firm's bargaining power may enable it to perform more loosely on credit lines without jeopardising its credit standing. We thus adjust the credit score for large firms with a discretized variable based on net sales (starting from 0 for firms with net sales below 100 million euro and up to 5 for firms with net sales above 5 billion euro), with an inverse effect on the probability of default. Our conjecture is supported by the model.

5.5.3 Integrated model

Following a standard approach in rating systems for the integration of information from different sources (e.g. Figini and Giudici, 2011, Giannozzi *et al.*, 2013, Balduini *et al.*, 2017), the partial scores obtained from the financial model and the credit behaviour model are integrated by means of a logistic regression that yields the final score; in turn, this is transformed into a PD via the inverse logit function (eq. 3).

The integration approach produces four distinct models by size (micro, small, medium, and large firms) according to the EC definition. This procedure reflects some underlying facts. We find that credit behaviour information has a large explanatory power for micro and small firms with a larger weight than that of the financial statement model. For medium and large firms, when an ordinary financial statement is available, the accounting module has a larger relative weight than that of the credit behaviour model. The ratio between the standardized beta of the credit behaviour component and that of the financial statement component ranges between 0.80 for large firms to 2.06 for micro firms (Fig. 16).

Figure 16. Ratio of standardized betas between credit behaviour model and financial statement model



As described in section 5.1, the model has been developed with the observations from 2015 to 2018 as the training sample and those for 2014 and 2019 as the test sample. Within the logit modelling framework, this implies that the (in-sample) average firm-level PD equals the average default rate recorded in the training sample. Considering that more recent data about the economic cycle in Italy are somewhat different from those of the years 2015-2018, in the statistical model we apply the so-called alpha-adjustment. We adjust the intercept α to match the default rate recorded over the longer period 2014-2021, with the following formula.

$$\alpha_{adjusted} = \ln\left(\frac{1-Def\ Rate_{15-18}}{Def\ Rate_{15-18}} \cdot \frac{Def\ Rate_{14-21}}{1-Def\ Rate_{14-21}}\right) \quad (5)$$

We apply the alpha-adjustment procedure for each sub-model by size. The default rate in the longer period 2014-2021 is lower than that in-sample. Hence the average PD of each statistical model is adjusted on the downside (Table 11).

Table 11 – Alpha-adjustment by size

Size	Default Rate (2014-21) (%)	Default Rate (2015-18) (%)	Difference (%)	Alpha-adjustment
Micro	4.36	4.84	-0.48	-0.11
Small	3.39	3.64	-0.25	-0.07
Medium	2.63	2.76	-0.13	-0.05
Large	2.39	2.66	-0.27	-0.11

6. Validation

The best practices in the field of credit risk assessment envisage the separation of validation from model development and rating production. Separation guarantees the independence and impartiality of validators. At the Bank of Italy the responsibility for rating production and model validation is assigned to two separate units.[18] The validation activity involves a one-off internal validation step as well as an ongoing monitoring process.

Internal validation aims at checking that the rating process is carried out consistently with the best methodological practices, thus providing adequately robust and efficient PD estimates. This activity is performed each time a new version of the model is developed or a significant innovation is introduced. The checks are applied to the whole model, both in the statistical component and in the expert system component.

The goal of performance monitoring is assessing the stability of the rating system over time and verifying the predictive ability of the model on a yearly basis.

Since ICAS is mainly aimed at PD estimation, internal validation does not include the assessment of the Loss Given Default (LGD) nor of the Exposure At Default (EAD).[19] The validation and monitoring of the PDs are mainly performed with benchmarking and backtesting.

Benchmarking techniques compare the PDs estimated by the model with those estimated by other models on the same sample of firms. These techniques typically involve the calculation of the statistical distance between the models. The main limitation of this approach is the small number of ratings usually obtained with other models, since ICAS assesses small and medium-sized enterprises that are seldom assessed with other systems. Besides, benchmarking requires strong confidence in the rating system used for the comparison.

Backtesting procedures are employed to compare the ratings estimated ex-ante with the number of defaults observed ex-post. The literature provides several statistical tests which help assess the model accuracy. Within the validation framework, three aspects are typically considered: discriminatory power, predictive power (or calibration quality), and performance stability.

Discriminatory power tests assess the ability of the model to distinguish firms on the basis of their future status (default or non-default) over a predefined time horizon. Predictive power tests compare the number of defaults that actually occurred in a given rating class with the number of defaults predicted by the model. Stability tests concern the ability of the rating system to distinguish between real causes/effects and purely random relationships; unstable rating systems have a disappointing performance (i.e. they lose discriminatory power or predictive power) if applied to databases other than that for which they were developed.

[18] This is in line with a specific ECB Governing Council recommendation. While approving the Bank of Italy's ICAS on 4 July and 7 November 2013, the Governing Council issued several recommendations, in particular that internal validation and monitoring of the model performance should be separated from methodology development and rating production.

[19] The approach that tackles also the analysis of the LGD and the EAD is typically followed in the IRB model validation (see also BIS, 2005).

Discrimination, prediction and stability of a model are not necessarily correlated. For instance, a rating system may show a high discriminatory power but a poor predictive performance; this can occur e.g. in the presence of small samples, as is often the case in practice. Alternatively, a rating system could have a good predictive power but a weak discriminatory power. This could occur if, for instance, in a population of obligors with an actual default rate equal to x per cent, the rating system assigns ratings randomly, attributing a probability of default of x per cent to each rating class; hence the probability of default is correctly assigned to each rating class, but the discriminatory power is nil.

Discriminatory power and predictive power are linked by an asymmetric relationship. Obtaining a good calibration (i.e. predictive power) is sometimes difficult; we note that it is easier to calibrate a model with a good discriminatory power than to improve the discriminatory power of the model, even if it is well calibrated. Furthermore, the discriminatory power of a rating system limits the quality of its calibration: it is the discriminatory power of a rating system that a priori makes it possible to obtain probabilities close to 0 or 1.

The next three subsections present the tools for discriminatory power analysis, predictive power analysis, and stability analysis, respectively. Section 7 shows the test results for our model.

6.1 Discriminatory power

The main purpose of a rating system is to distinguish between 'sick' and 'healthy' units (firms), depending on whether or not the occurrence of the default event is considered likely for each of them within a given time horizon. In most cases, the model is expected to draw a line between the two types of units; the most common procedure involves setting a cut-off probability. The units that have an estimated default probability below (above) the cut-off level are considered healthy (sick). The model must have discriminatory power, that is it should show precision in assigning a default probability below (above) the cut-off level to the healthy (sick) firms.

The discriminatory power should reflect the following properties:

- specificity, that is the ability to correctly classify the units for which the event does not occur;
- sensitivity, that is the ability to correctly classify the units for which the event occurs.

There is a clear trade-off between specificity and sensitivity. For instance, as the cut-off level for the probability of default increases, the model will be less specific but more sensitive.

The standard measure for the discriminatory power of a model is the Receiver operating characteristic (ROC) curve, which in turn yields the AUROC statistic (see section 4.2). We note that the evidence about the units gone into default and those that survived in the time period under consideration is only one of the possible realizations from the probability distributions of the defaulters and the non-defaulters. In other words, the default phenomenon has a stochastic nature. If the same experiment were repeated, different realizations of these distributions would be obtained, with every realization showing a different value of the AUROC.

The stochastic nature of the default phenomenon can be investigated with the statistic known as U of Mann-Whitney, which is strictly linked to AUROC (Cortes and Mohri, 2005); the statistical properties of the Mann-Whitney U are thus applicable to the study of the stochastic behaviour of the AUROC.

The Mann-Whitney U statistic can help quantify how far the value achieved by the AUROC is from its expected value, by calculating appropriate confidence intervals; this allows for hypothesis testing, for example to check whether the discriminatory power of the rating system is significantly different from the value under total randomness.

6.2 Predictive power

The predictive power (or calibration quality) refers to the model's ability to identify the 'real' probability of default for an individual debtor or a class of homogeneous debtors. Since the real PD is unknown, it is not possible to accurately estimate it. Therefore, to test the quality of the PD calibration the observed (ex-post) default frequency is compared with the estimated (ex-ante) probability of default and hypothesis testing is performed. Typically, a rating system includes various classes of risk and the validator must assess each risk class separately.

Conditional and unconditional tests can be used. The notions of conditional and unconditional testing can be best understood if related to the notions of 'point-in-time' (PIT) and 'through-the-cycle' (TTC) probability of default. The probability of default can be based (or 'conditioned') on the current state of the economy, for example by including macroeconomic variables in the regression equation. Typical macro variables are the GDP growth rate and the unemployment rate. In this case the estimated probabilities of default are PIT PDs. Credit events can then be treated as independent, because most of the dependence among them is captured by incorporating the macroeconomic variables into the PD forecasts. In contrast, the unconditional estimates of the PD are not based on the current state of the economy. Such estimates, based on data referring to a complete economic cycle, are denominated TTC PDs. When using unconditional PDs, no independence assumption between the credit events can be accepted, given that the change in the observed default rates cannot any longer be explained by the change in the conditional PDs, which are themselves random variables (Tasche, 2006).

The estimated PDs provided by ICAS are generally of the PIT type (Auria *et al.*, 2021) because they capture the perspective creditworthiness of firms as closely as possible, to better control the credit risk of collateral in monetary policy refinancing operations. Thus, the validation of the predictive power of ICAS is performed mainly with conditional tests on the state of the economy. These tests are illustrated below.

The first approach for testing multiple risk classes at the same time involves multiple comparisons. As an alternative, a single statistical test is used to compare risk classes simultaneously (joint testing). When multiple testing is applied, each risk class is assessed individually in the first stage. Under the usual assumption of independence among the defaults, the number of defaults by risk class follows a binomial distribution. In the second stage the results of the individual tests are considered, to check if the rating system adopts appropriate PD estimates for all classes.

It is important to clarify a definition issue. The ICAS 'statistical model' produces probabilities; based on them, each sample company is assigned to a specific risk class. To assess the eligibility of the company

and obtain the prudential haircut for monetary policy purposes, the ICAS 'rating system' gives each company a probability equal to that of the upper limit of the risk class.

For the validation of ICAS, we perform multiple testing to check: 1) for the absence of underestimation of the ICAS *system* PDs in each rating class (adopting a risk aversion profile towards default typical of a 'supervisor'); 2) whether the ICAS *system* is too conservative or too loose, i.e. whether its PDs are in line with actual defaults (adopting a risk aversion profile of a 'production unit' with business objectives); 3) whether the ICAS *model* PDs are in line with the actual defaults.

Under the null hypothesis that 'the expected defaults are higher than those observed' (the supervisor view), a one-tailed binomial test is conducted for each rating class. The test distribution has an average value equal to the PD value of the upper bound in each rating class. The test is evaluated at the 99% confidence level. When the number of actual defaults is significantly lower than that of the expected defaults for all rating classes, the ICAS system is prudential, i.e. it does not lead to underestimating default risk.

Employing as the null hypothesis the statement 'the expected defaults are equal to those observed, for each rating class' (the production unit view) – which is equivalent to 'the ICAS system is precise' – a two-tailed binomial test is conducted for each rating class. As in the previous case, the distribution of each test has the average value equal to the upper end of the probability range of each rating class. The test is evaluated at the 99% confidence level. If the number of actual defaults is significantly higher or lower than that of the expected defaults for a given rating class, the null hypothesis is rejected.

Lastly, under the null hypothesis 'the expected defaults of the statistical model underlying the ICAS rating system are equal to those observed, for each rating class' – which is equivalent to 'the ICAS model is precise' – a two-tailed binomial test is conducted for each rating class. The distribution of each test has the average value equal to the average of the probabilities of default assigned by the statistical model to the borrowers that belong to each rating class. The test is evaluated at the 99% confidence level. If the number of actual defaults is significantly higher or lower than that of the expected defaults for some rating class, the hypothesis 'the ICAS model is precise' is rejected.

However, when multiple testing analysis is conducted, the so-called alpha-inflation problem should be accounted for. This is due to the fact that, as the number of hypotheses to be simultaneously tested increases, the probability of type 1 error on the global hypothesis rises.

Some methods can be employed to keep this error under control. The most common in the literature consists in increasing the *p*-values on the single class tests. Some of these methods are applied also in the ICAS validation process, either under the supervisor approach or under the production unit approach.

Within the family of joint tests, those put forward by Spiegelhalter and Hosmer and Lemeshow are the most commonly used. The Spiegelhalter test, based on the Brier Score,[20] combines the information on the

[20] The Brier Score is a score function that measures the accuracy of probabilistic predictions. For unidimensional predictions, it is strictly equivalent to the mean squared error as applied to predicted probabilities. The Brier score is applicable to tasks in which predictions must assign probabilities to a set of mutually exclusive discrete outcomes or classes. The set of possible outcomes can either be binary or categorical in nature, and the probabilities assigned to this set of outcomes sum to one (where each individual probability is in the range of 0 to 1). The Brier score can be thought of as a cost function. More precisely, across all items $i \in 1...N$ in a set of N predictions, the Brier score measures the

calibration quality available at the individual level. Under the null hypothesis of perfect calibration (the expected value of the default distribution is equal to the PD estimate for each unit), the test statistic is distributed as a standardized normal. The test is based on the assumption of independence among the defaults. The Spiegelhalter test results should be interpreted carefully, since the test statistic is based on a weighted average PD that is estimated on the entire sample of debtors. Only an average overestimation/underestimation of the probabilities of default causes the rejection of the null hypothesis of perfect calibration at the individual level. The test therefore fails to identify as unacceptable those situations in which the overestimation of PDs for some units is offset by the underestimation of PDs for other units in the sample. To overcome this issue, the Hosmer-Lemeshow test compares expected defaults and realized defaults within each rating class. The test, like the Spiegelhalter test, uses a normal approximation of the binomial distribution that leads to a chi-square distribution of the statistic.[21]

6.3 Stability

The stability of a rating system is usually validated by applying the rating system to a data sample different from the one which is used to develop the system. The latter is referred to as the 'training sample', while the sample employed for stability evaluation is called the 'validation sample'. There are three ways to construct the validation sample:

- out-of-sample: the training sample is obtained by randomly selecting a part of a larger sample. The complement of the training sample within the larger sample is used for validation;
- out-of-universe: the training sample is formed by deterministically selecting a part of a larger sample (e.g. considering only data from specific regions or sectors). The complement of the training sample within the larger sample is used for validation;
- out-of-time: the training sample is made up of data covering a given time period. Data from outside this period is used for validation.

In practice, the availability of data may be an issue. The full starting sample, to be split into a training sample and a validation sample, is often small. If, on the one hand, the training sample is too small, there will be uncertainty in the estimation and an excessive adaptation to the available data (over-fitting); then the validation will end up confirming what is obvious. If, on the other hand, the validation sample is too small, validation will be of little use.

An approach that is often applied in practice is the 'walk-forward testing'. The rating system to be validated is set up using data from a certain period of time only. Data from the following periods are added gradually and the exercise is repeated, leading to a time series of results that can be used to assess performance. This approach is very useful because it replicates the process which the rating system will undergo over time. In the light of new data, there will be an ongoing review and re-fitting of the system.

mean squared difference between the predicted probability assigned to the possible outcomes for item i and the actual outcome o_i. Therefore, the lower the Brier score is for a set of predictions, the better the predictions are calibrated.

[21] This approximation could be questionable with a small number of borrowers.

7. Model performance: backtesting and comparison with the previous model

In this section we illustrate the performance of the model in terms of discriminatory power, predictive power, and stability by means of a backtesting exercise. Besides, we benchmark the model performance against that of the previous version of the model (Giovannelli et al, 2020). Backtesting is based on the following criteria:

- discriminatory power - the model's AUROC should be as large as possible (above 70 per cent);
- predictive power - in the first place, the model's default forecast should be conservative; besides, a model that provides precision, by forecasting default rates as close as possible to actual default rates, should be preferred;
- stability - the measures of discriminatory power and predictive power should change as little as possible over time; validation should favour the model with a stationary behaviour.

We recall that:

- the previous model was estimated on the defaults of the years 2015-2016, with an alpha correction based on 2016;
- the new model is estimated on the defaults of the years 2015-2018, with an alpha correction based on 2014-2021.

A fully informative validation would in principle involve an out-of-sample analysis. We employ the actual defaults for the years 2019-2021, available at the time of the analysis. This procedure is partially in-sample, because the alpha-adjustment also employs the defaults of the years 2019-2021. However, we also note that the overlap between the estimation sample and the test sample is limited, and alpha is calibrated over a suitably large number of years.

The performance of the two models is examined at the following reference dates: December 2018 (default period from January to December 2019), June 2019 (default period July 2019-June 2020), December 2019 (default period January-December 2020), June 2020 (default period July 2020- June 2021), December 2020 (default period January-December 2021). For completeness, the probabilities of default have been compared with the subsequent defaults of fractional type and binary type.

In section 5 we describe the time period used for the model development carried out by the CRA Division. In this section we report the independent validation of the model carried out by the FRC Division. The validation function also looks at different default periods compared to those used in the estimation process. The validation analysis refers to the one-year default period (all PDs have a one-year horizon).

For completeness, the PDs are validated using both the binary and the fractional default definition (section 5.2).

In terms of discriminatory power, the new model achieves fairly large AUROC values, at 84 per cent or better. These are always larger that the corresponding figures for the previous model at each date (Table 12). For large firms the AUROC is as large as 90 per cent (see Appendix 2).

Table 12 – AUROC

(percentage values)

Default period	December 2018	June 2019	December 2019	June 2019	December 2020
Binary defaults					
Previous model	87.6	87.0	87.6	87.4	83.3
New model	88.5	88.1	88.9	88.6	84.5
Fractional defaults					
Previous model	83.3	84.4	84.3	83.8	81.6
New model	84.0	84.8	84.9	84.3	82.8

In terms of predictive power, the two models pass the test on the absence of underestimation of the PD in each rating class. As concerns accuracy, the new model generally performs better than the previous one. In fact, using the third test presented in the previous section and considering the number of rating classes for which the relevant model presents a significant deviation between the number of actual defaults and the number of estimated defaults, the new model features the absence of 'slack' classes, that is in no case is the number of forecast defaults smaller than that of the actual ones.

Table 13 – Number of rating classes: estimated and actual defaults

Default period	December 2018	June 2019	December 2019	June 2019	December 2020
Binary defaults					
Previous model	2	1	1	1+4	1
New model	2	3	5	4	1
Fractional defaults					
Previous model	1	0	2	4	0
New model	2	4	5	5	2

Note: the total number of rating classe (CQS) is eight. The numbers

- in blue show for how many classes: forecasted defaults > effective defaults (the model is conservative);
- in red show for how many classes: forecasted defaults < effective defaults (the model is slack);
- in black indicate a situation in which for every class: forecasted defaults = effective defaults (the model is precise).

The standard deviation of the results presented in the above tables turns out to be smaller for the new model in almost all cases. Thus, the new model exhibits a lower volatility of the results.

We can conclude that, compared to the previous version of the statistical model, the new model features a larger discriminatory power, a larger predictive power, and greater stability. In addition, as argued above, the new model is more transparent for analysts. All these properties are highly desirable and represent a substantial improvement over the previous model.

Next, we examine the behaviour of the in-sample calibration version of the model (or base model), to find out if and to what extent it provides robust PD estimates with respect to the actual defaults, without considering the alpha-adjustment.

First, we note that the discriminatory power of the base model is the same as reported in Table 12. A shift of the PDs such as that implied by the alpha-adjustment does not affect the ordering of the PDs among themselves.

The predictive power worsens, since in this case there is a larger number of 'not precise' classes, even if the base model is never slack, as it happens for the previous model (Table 14). That is, the base model is rather conservative; based on the data, the alpha-adjustment partially reduces the degree of prudence of the model.

Table 14 – Number of rating classes: comparison between estimated and actual defaults
(base model)

Default period	December 2018	June 2019	December 2019	June 2019	December 2020
Binary defaults					
Previous model	2	1	1	1+4	1
New model	5	4	5	5	2
Fractional defaults					
Previous model	1	0	2	4	0
New model	4	5	5	5	5

Note: the total number of rating classe (CQS) is eight. The numbers
- in blue show for how many classes: forecasted defaults > effective defaults (the model is conservative);
- in red show for how many classes: forecasted defaults < effective defaults (the model is slack);
- in black indicate a situation in which for every class: forecasted defaults = effective defaults (the model is precise).

As concerns the stability of the results, the base part of the statistical model exhibits a less volatile behaviour than that of the previous model, as well as that of the recalibrated model. The standard deviation of the results presented in the previous tables is almost always smaller than in the base model.

To conclude, the statistical model described in this paper has a robust structure, as it yields PD estimates that enable the system to discriminate between healthy and sick firms and that never predict a number of defaults below that of actual defaults. Likewise, it appears that the base model is highly prudent. This property can be managed by means of the alpha-adjustment (see section 5.5.3).

The validation analysis was also conducted for each size class of the firms for which a PD is provided by the model. This approach has confirmed the evidence regarding similarities and differences between the new model and the previous one. We also find that the discriminatory power of the statistical model improves as the size of the company increases. This is a favourable property, as the largest exposures are usually held towards the largest firms.

As far as predictive power is concerned, although the new model is more prudent than the previous version towards smaller firms, the two models tend to converge and become more precise as the size of the company increases (see Appendix 2). This result will guide possible future methodological developments of the model.

8. Conclusions

This paper describes the methodology underlying the Bank of Italy's statistical credit assessment model and its validation process. Each month the model produces with an automated procedure an estimate of the one-year PD for 370,000 Italian non-financial firms. The model preserves the transparency of the traditional logistic model, while trying to improve the performance with ML techniques that capture complex relationships of some variables with the default event.

Empirically, the model shows satisfactory features that enable it to properly discriminate between healthy and risky firms, with an AUROC statistic between 84 and 90 per cent. The stability of the results is fairly high. Furthermore, the discriminatory power of the model improves as the size of the company increases, thus ensuring a proper evaluation of the largest exposures in monetary policy operations.

The architecture of the model will enable us to work on possible future improvements, such as the integration of macroeconomic variables in the calibration of the alpha-adjustment to better capture the average default rate dynamics, as well the inclusion of other information, such as the analysis of climate risk and press news, to further improve the discriminatory power of the model. As concerns the econometric approach, the introduction of panel logistic techniques could improve the consistency of the parameters' estimation.

References

Alberici A., 1975, "Analisi dei bilanci e previsione delle insolvenze", *Isedi*, Milano.

Altman E.I., 1968, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, 23, 589-609.

Altman E.I.,1983, "Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptcy", *Wiley Interscience*, John Wiley and Sons.

Altman E.I., Halderman R.G., Narayanan P., 1977, "Zeta-analysis. A new Model to Identify Bankruptcy Risk of Corporations", *Journal of Banking and Finance*, 1, 29-54.

Altman E.I., Iwanicz-Drozdowska, M., Laitinen, E. K. and Suvas, A., 2017, "Financial Distress Prediction in an International Context: A Review and Empirical Analysis of Altman's Z-Score Model", *Journal of International Financial Management and Accounting*, 28, 131-171.

Altman E.I., Sabato G., 2007, "Modelling Credit Risk for SMEs: Evidence from the U.S. market", *ABACUS*, 43(3).

Altman E.I., Sabato G., Wilson L., 2010, "The Value of Non-financial Information in SME Risk Management", *The Journal of Credit Risk*, 6(2), 1-33.

Angori G., Aristei D., Gallo M., 2020, "Banking Relationship, Firm-size Heterogeneity and Access to Credit: Evidence from European Firms", *Finance Research Letters*, 33, March 2020.

Antunes A., Gonçalves H., Prego P., 2016, "Firm Default Probabilities Revisited", *Economic Studies*, Banco de Portugal, April 2016.

Auria L., Bingmer M., Mateo C., Graciano C., Charavel C., Gavilá S., Iannamorelli A., Levy A., Maldonado A., Resch F., Rossi A.M., Sauer S., 2021, "Overview of Central Banks' In-house Credit Assessment Systems in the Euro area", *Occasional Paper Series*, ECB, 284.

Balduini A., Dwyer D., Gianfreda S., Hemminki R., Yang L., Zhao J.Y., 2017, "Combining Financial and Behavioral Information to Predict Default for Small and Medium-Sized Enterprises – A Dynamic Weighting Approach", Moody's Analitics and Credit Data Research.

Barboza F., Kimura, H., Altman, E., 2017, "Machine Learning Models and Bankruptcy Prediction", *Expert Systems with Applications*, 83, 405-417.

BIS, 2005, "Studies on the Validation of Internal Rating Systems", *Working Paper*, 14, https://www.bis.org/publ/bcbs_wp14.pdf.

Blum M., 1974, "Failing Company Discriminant Analysis", *Journal of Accounting Research*, 12(1), 1-25.

Bonaccorsi di Patti E., D Ignazio A., Gallo M, Micucci G., 2015, "The Role of Leverage in Firm Solvency: Evidence from Bank Loans", *Italian Economic Journal,* 1(2), 253-286, July 2015.

Brier, G.W., 1950, "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review*, 78(1), 1-3.

Campbell J.Y., Hilscher J., Szilagyi J., 2008, "In Search of Distress Risk", *The Journal of Finance*, 63(6), 2899-2939.

Carmignani A., Omiccioli M., 2007, "Costs and benefits of creditor concentration: an empirical approach", *Working Papers (Temi di discussione)*, 645, Bank of Italy.

Cascarino, G., Moscatelli, M., Parlapiano, F., 2022, "Explainable artificial intelligence: Interpreting default forecasting models based on Machine Learning", *Occasional Paper (Questioni di Economia e Finanza)*, 674, Banca d'Italia.

Charalambakis C., Garrett I., 2019, "On corporate financial distress prediction: what can we learn from private firms in a developing economy? Evidence from Greece", *Review of quantitative finance and accounting*, 52, 467-491.

Chawla N.V., 2009, "Data Mining for Imbalanced Datasets: an Overview", in: Maimon O., Rokach L. (eds) *Data mining and knowledge discovery handbook*, 875-886, Springer, Boston, MA.

Chen K.H., Shimerda T.A., 1981, "An Empirical Analysis of Useful Financial Ratio", *Financial Management*, 10(1), 51-60.

Cortes, C., Mohri, M., 2005, "Confidence Intervals for the Area under the ROC Curve", *Advances in neural information processing systems*, 305-312.

Deakin E., 1972, "A Discriminant Analysis of Predictors of Business Failure", *Journal of Accounting Research*, 10(1), Spring, 167-179.

De Socio A., Michelangeli V., 2017, "Modelling Italian Firms' Financial Vulnerability", *Journal of Policy Modeling*, 39, 147-168.

Dumitrescu E., Hué S., Hurlin C., Tokpavi S., 2022, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects", *European Journal of Operational Research*, 297(3), 1178-1192.

EBA, 2021, "EBA Discussion Paper on Machine Learning for IRB Models", *EBA/DP/2021/04*, November 2021.

Edmister R., 1972, "An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction", *Journal of Financial and Quantitative Analysis*, 7(2), 1477-1493.

Fantazzini D., Figini S., 2009, "Random Survival Forests Models for SME Credit Risk Measurement", *Methodology and Computing in Applied Probability*, 11, 29-45.

Fawcett T., 2004, "ROC graphs: Notes and Practical Considerations for Researchers", *Machine Learning*, 31(1), 1-38.

Figini S., Giudici P., 2011, "Statistical Merging of Rating Models", *The Journal of the Operational Research Society*, 62(6), 1067-1074.

FSB, 2010, "Principles for Reducing Reliance on CRA Ratings".

Gallucci C., Santulli R., Modina M., Formisano F., 2022, "Financial Ratios, Corporate Governance and Bank-firm Information: a Bayesian Approach to Predict SMEs' Default", *Journal of Management and Governance*, February 2022, https://doi.org/10.1007/s10997-021-09614-5

Gentry J.A., Newbold P., Whitford D.T., 1985, "Classifying Bankrupt Firms with Funds Flow Components", *Journal of Accounting Research*, 23(1), Spring, 146-160.

Giannozzi A., Altman E.I., Roggi O., Sabato G., 2013, "Building SME Rating: is it Necessary for Lenders to Monitor Financial Statements of the Borrowers?", *Bancaria Editrice*, 10, 54-71.

Giovannelli F., Iannamorelli A., Levy A., Orlandi M., 2020, "The in-house credit assessment system of Banca d'Italia", *Occasional Papers*, 586, Bank of Italy.

Giovannelli F., Iannamorelli A., Levy A., Orlandi M., 2023, "The Bank of Italy's In-House Credit Assessment System for Non-financial Firms", in: Scalia, A. (ed.) *Financial Risk Management and Climate Change Risk. The Experience in a Central Bank*, Springer, 107-137.

Grandia R., Hänling P., Lo Russo M., Âberg P., 2019, "Availability of high-quality Liquid Assets and Monetary Policy Operations: an Analysis for the Euro Area", *Occasional Paper Series*, ECB, 218.

Hajek P., 2012, "Credit Rating Analysis Using Adaptive Fuzzy Rule-based Systems: an Industry-specific Approach", *Central European Journal of Operations Research*, 20(3), 421-434.

Jiménez G., Lopez J., Saurina J., 2009, "Empirical Analysis of Corporate Credit Lines", *The Review of Financial Studies*, 22, 5069-5098.

Kosekova K., Maddaloni A., Papoutsi M., Schivardi F., 2023, "Firm-Bank Relationship: a Cross-country Comparison", *Working Paper Series*, ECB, 2826.

Lee S., Choi W.S., 2013, "A Multi-Industry Bankruptcy Prediction Model Using Back-propagation Neural Network and Multivariate Discriminant Analysis", *Expert System with Application*, 40(8), 2941-2946.

Luerti A., 1992, "La previsione dello stato di insolvenza delle imprese. Il modello AL/93", *Etaslibri*, Milano.

Micha B., 1984, "Analysis of Business Failure in France", *Journal of Banking and Finance*, 8, 281-291.

Modigliani F., Miller M., 1958, "The Cost of Capital, Corporation Finance and the Theory of Investment", *American Economic Review*, 48(3), 261-297.

Moscatelli M., Narizzano S., Parlapiano F., Viggiano G., 2020, "Corporate Default Forecasting with Machine Learning", *Expert Systems with Application*, 161, 113567.

Mossman C.E., Bell G.G., Swartz L.H., Turtle H., 1998, "An Empirical Comparison of Bankruptcy Models", *American Economic Review*, 48(2), 35-54.

Norden L., Weber M., 2010, "Credit Line Usage, Checking Account Activity, and Default Risk of Bank Borrowers", *The Review of Financial Studies*, 23, 3665-3699.

Ohlson J., 1980, "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, 18(1), Spring, 109-131.

Ortiz-Molina H., Penas M.F., 2008, "Lending to Small Business: the Role of Loan Maturity in Addressing Information Problems", *Small Business Economics*, 30(4), 361-283.

Partridge J., Yang H., LaSalvia T., 2017, “Machine Learning: Challenges, Lessons, and Opportunities in Credit Risk Modeling”, Moody’s, 1 July.

Tasche D., 2006, “Validation of Internal Rating Systems and PD Estimates”, *The Analytics of Risk Model Validation*, 169-196.

Van Gestel T., Baesens B., Van Dijcke P., Suykens J.A.K., Crispiniano Garcia, J.B., Alderweireld T., 2005, “Linear and Non-Linear Credit Scoring by Combining Logistic Regression and Support Vector Machines”, *Journal of Credit Risk*, 1(4), Fall 2005.

Zhao Y.J., Dwyer D.W., Zhang J., 2014, “Usage and Exposures at Default of Corporate Credit Lines: an Empirical Study”, *Journal of Credit Risk*, 10(1).

Appendix 1. Model development

The first step in model development consists in (i) the choice of the dependent variable with the specification of the default definition, and (ii) the collection of the input data with an initial analysis on quality and consistency.

The second step consists in splitting the data sample for the purpose of estimation and testing (see section 5.1).

The third step, for each component of the model (11 financial statement sub-models and 3 credit behaviour sub-models), is a univariate analysis on a list of variables of interest that are potentially predictive of default. The variables are chosen after consulting credit analysts and reviewing the literature and other credit scoring models. For both the financial statement sub-models and the credit behaviour sub-models, we divide the list of variables into classes that reflect different hypotheses regarding a firm's creditworthiness (profitability, debt sustainability, liquidity, etc.). Within the univariate analysis, we analyse and treat the outliers that could cause distortions in the development of the models. For continuous variables, we examine the frequency distribution and the key statistics and apply winsorization methods to the outliers, by replacing each outlier with a given percentile of the frequency distribution (Barnett and Lewis, 1994). Besides, we examine missing or inconsistent values for each variable. When a missing or inconsistent value is related to a situation that could be 'extremely positive' or 'extremely negative', we replace that value with the extreme value used in the winsorization process. When a missing value is not related to a positive or negative condition, it can be replaced with the median or average value of the distribution or with a dummy variable. For example, the ratio drawn amount to granted amount for a certain type of credit line reported in the NCR could be missing. In this case we assign a value of 0 (positive value) and introduce a dummy variable to set the credit score equal to the average default rate of the firms that do not have this type of credit line.

The aim of the univariate analysis is to shorten the list of variables and to identify, for each group of candidate regressors, the most significant ones for the multivariate analysis. The variable selection uses the following criteria:

- variables with a lot of missing values and outliers are dropped;
- variables not having a linear and monotonic relationship with default are usually dropped. Nevertheless, for some variables of interest (such as sales growth and days of receivables turnover) we consider as potential predictors also the discretized transformation obtained with the application of decision tree techniques;
- using a *t*-test, the variables with insignificant differences in the mean value between the default and non-default groups are dropped;
- using univariate logit regression for the probability of default, the variables with an accuracy ratio below 55 per cent are dropped;
- for each risk group, we retain the variables with a high accuracy ratio and some other variables based on a qualitative assessment.

The fourth step involves the analysis of collinearity. To assess the possible informative redundancies and reduce the collinearity between variables, we verify their independence. To this aim, we construct a correlation matrix between the variables selected in the previous step. If two or more variables have a correlation above 0.7 or below -0.7 we keep the one with the highest discriminatory power.

For each sub-model, the final step is the selection process that involves the choice of the final combination of variables and the definition of their relative weights. We use a multivariate logistic regression with a statistical selection of indicators performed via a stepwise algorithm with a 1% minimum significance level for a variable to enter the model. We review the results of the statistical process to make sure that a) all the relevant risk areas are represented, and b) the estimated sign of the coefficient of each indicator is consistent with the underlying economic assumptions.

Appendix 2. Firm size

The behaviour of the statistical model against the previous model can be further investigated by splitting the sample according to firm size.

The new model displays a higher discriminatory power. This is even more true in the case of large firms, both under the definition of binary default and of fractional default. Larger firms generally have larger credit exposures, hence the ability to discriminate correctly in this context can be very important. These results are shown in the following table.

Table A2.1 – AUROC by firm size
(percentage values)

Default period	December 2018	June 2019	December 2019	June 2019	December 2020
Binary defaults					
Micro firms					
Previous model	87.1	86.6	87.0	87.3	82.9
New model	87.6	87.3	88.0	88.1	83.5
Small firms					
Previous model	88.3	87.7	88.7	87.6	84.1
New model	90.2	89.5	90.4	89.5	86.3
Medium firms					
Previous model	88.2	87.0	88.3	86.1	84.3
New model	90.7	89.2	90.6	89.1	86.4
Large firms					
Previous model	89.3	85.0	85.7	84.5	84.1
New model	90.9	86.8	88.4	89.0	90.0
Fractional defaults					
Micro firms					
Previous model	82.5	83.9	83.5	83.2	80.6
New model	82.7	83.8	83.8	83.3	81.2
Small firms					
Previous model	84.2	84.8	82.9	83.9	82.9
New model	86.2	86.4	86.6	85.6	85.5
Medium firms					
Previous model	84.5	83.8	85.8	83.6	82.6
New model	86.7	85.5	87.1	86.3	85.3
Large firms					
Previous model	86.7	82.9	83.5	79.7	78.5
New model	87.3	85.5	86.2	82.7	84.9

The new version of the model is more prudent than the previous one, in particular for the micro enterprises, which are approximately 70 per cent of all sample firms. As the size increases, the two models become more precise. The previous model shows a slack in several periods, under both binary and fractional defaults, while the new model never presents this problem. This result is reported in the following table.

Table A2.2 – Number of classes with issues on estimated PDs, by firm size

Default period	December 2018	June 2019	December 2019	June 2019	December 2020
Binary defaults					
Micro firms					
Previous model	1	0	1	4	0
New model	2	3	4	4	1
Small firms					
Previous model	2	0	0	3	1
New model	0	0	1	3	0
Medium firms					
Previous model	0	1	0	0	0
New model	1	0	2	2	0
Large firms					
Previous model	0	0	1	0	0
New model	0	0	0	0	0
Fractional defaults					
Micro firms					
Previous model	1	0	3	4	2
New model	2	3	5	5	1
Small firms					
Previous model	0	0	0	4	2
New model	0	0	4	4	1
Medium firms					
Previous model	0	0	0	2	1
New model	1	0	2	2	2
Large firms					
Previous model	0	1	0	0	0
New model	0	0	0	0	0

Note: the total number of rating classes (CQS) is eight. The numbers

- in blue show for how many classes: forecasted defaults > effective defaults (the model is conservative);
- in red show for how many classes: forecasted defaults < effective defaults (the model is slack);
- in black indicate a situation in which for every class: forecasted defaults = effective defaults (the model is precise).

Recently published papers in the 'Markets, Infrastructures, Payment Systems' series

n. 16 Cross-Currency Settlement of Instant Payments in a Multi-Currency Clearing and Settlement Mechanism, *by Massimiliano Renzetti*, *Fabrizio Dinacci and Ann Börestam* (Research Papers)

n. 17 What's ahead for euro money market benchmarks?, *by Daniela Della Gatta* (Institutional Issues) (in Italian)

n. 18 Cyber resilience per la continuità di servizio del sistema finanziario, *by Boris Giannetto and Antonino Fazio* (Institutional Issues) (in Italian)

n. 19 Cross-Currency Settlement of Instant Payments in a Cross-Platform Context: a Proof of Concept, *by Massimiliano Renzetti, Andrea Dimartina, Riccardo Mancini, Giovanni Sabelli, Francesco Di Stasio, Carlo Palmers, Faisal Alhijawi, Erol Kaya, Christophe Piccarelle, Stuart Butler, Jwallant Vasani, Giancarlo Esposito, Alberto Tiberino and Manfredi Caracausi* (Research Papers)

n. 20 Flash crashes on sovereign bond markets – EU evidence, *by Antoine Bouveret, Martin Haferkorn, Gaetano Marseglia and Onofrio Panzarino* (Research Papers)

n. 21 Report on the payment attitudes of consumers in Italy: results from ECB surveys, *by Gabriele Coletti, Alberto Di Iorio, Emanuele Pimpini and Giorgia Rocco* (Institutional Issues)

n. 22 When financial innovation and sustainable finance meet: Sustainability-Linked Bonds, *by Paola Antilici, Gianluca Mosconi and Luigi Russo* (Institutional Issues) (in Italian)

n. 23 Business models and pricing strategies in the market for ATM withdrawals, *by Guerino Ardizzi and Massimiliano Cologgi* (Research Papers)

n. 24 Press news and social media in credit risk assessment: the experience of Banca d'Italia's In-house Credit Assessment System, *by Giulio Gariano and Gianluca Viggiano* (Research Papers)

n. 25 The bonfire of banknotes, *by Michele Manna* (Research Papers)

n. 26 Integrating DLTs with market infrastructures: analysis and proof-of-concept for secure DvP between TIPS and DLT platforms, *by Rosario La Rocca, Riccardo Mancini, Marco Benedetti, Matteo Caruso, Stefano Cossu, Giuseppe Galano, Simone Mancini, Gabriele Marcelli, Piero Martella, Matteo Nardelli and Ciro Oliviero* (Research Papers)

n. 27 Statistical and forecasting use of electronic payment transactions: collaboration between Bank of Italy and Istat, *by* Guerino Ardizzi and Alessandra Righi (Institutional Issues) (in Italian)

n. 28 TIPS: a zero-downtime platform powered *by* automation, by *Gianluca Caricato, Marco Capotosto, Silvio Orsini and Pietro Tiberi* (Research Papers)

n. 29 TARGET2 analytical tools for regulatory compliance, *by Marc Glowka, Alexander Müller, Livia Polo Friz, Sara Testi, Massimo Valentini and Stefano Vespucci* (Institutional Issues)

n. 30 The security of retail payment instruments: evidence from supervisory data, *by Massimiliano Cologgi* (Research Papers)

n. 31 Open Banking in the payment system: infrastructural evolution, innovation and security, supervisory and oversight practices, *by Roberto Pellitteri, Ravenio Parrini, Carlo Cafarotti and Benedetto Andrea De Vendictis* (Institutional Issues) (in Italian)

n. 32 Banks' liquidity transformation rate: determinants and impact on lending, *by Raffaele Lenzi, Stefano Nobili, Filippo Perazzoli and Rosario Romeo* (Research Papers)

n. 33 Investor behavior under market stress: evidence from the Italian sovereign bond market, *by Onofrio Panzarino* (Research Papers)

n. 34 Siamese neural networks for detecting banknote printing defects, *by Katia Boria, Andrea Luciani, Sabina Marchetti and Marco Viticoli* (Research Papers) (in Italian)

n. 35 Quantum safe payment systems, *by Elena Bucciol and Pietro Tiberi*

n. 36 Investigating the determinants of corporate bond credit spreads in the euro area, *by Simone Letta and Pasquale Mirante*

n. 37 Smart Derivative Contracts in DatalogMTL, *by Andrea Colombo, Luigi Bellomarini, Stefano Ceri and Eleonora Laurenza*

n. 38 Making it through the (crypto) winter: facts, figures and policy issues, *by Guerino Ardizzi, Marco Bevilacqua, Emanuela Cerrato and Alberto Di Iorio*

n. 39 The Emissions Trading System of the European Union (EU ETS), *by Mauro Bufano, Fabio Capasso, Johnny Di Giampaolo and Nicola Pellegrini* (in Italian)

n. 40 Banknote migration and the estimation of circulation in euro area countries: the italian case, *by Claudio Doria, Gianluca Maddaloni, Giuseppina Marocchi, Ferdinando Sasso, Luca Serrai and Simonetta Zappa* (in Italian)

n. 41 Assessing credit risk sensitivity to climate and energy shocks, *by Stefano Di Virgilio, Ivan Faiella, Alessandro Mistretta and Simone Narizzano*

n. 42 Report on the payment attitudes of consumers in italy: results from the ecb space 2022 survey, *by Gabriele Coletti, Alberto Di Iorio, Emanuele Pimpini and Giorgia Rocco*

n. 43 A service architecture for an enhanced Cyber Threat Intelligence capability and its value for the cyber resilience of Financial Market Infrastructures, *by Giuseppe Amato, Simone Ciccarone, Pasquale Digregorio and Giuseppe Natalucci*

n. 44 Fine-tuning large language models for financial markets via ontological reasoning, *by Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili and Emanuel Sallinger*

n. 45 Sustainability at shareholder meetings in France, Germany and Italy, *by Tiziana De Stefano, Giuseppe Buscemi and Marco Fanari* (in Italian)

n. 46 Money market rate stabilization systems over the last 20 years: the role of the minimum reserve requirement, *by Patrizia Ceccacci, Barbara Mazzetta, Stefano Nobili, Filippo Perazzoli and Mattia Persico*

n. 47 Technology providers in the payment sector: market and regulatory developments, *by Emanuela Cerrato, Enrica Detto, Daniele Natalizi, Federico Semorile and Fabio Zuffranieri*

n. 48 The fundamental role of the repo market and central clearing, *by Cristina Di Luigi, Antonio Perrella and Alessio Ruggieri*

n. 49 From Public to Internal Capital Markets: The Effects of Affiliated IPOs on Group Firms, *by Luana Zaccaria, Simone Narizzano, Francesco Savino and Antonio Scalia*

n. 50 Byzantine Fault Tolerant consensus with confidential quorum certificate for a Central Bank DLT, *by Marco Benedetti, Francesco De Sclavis, Marco Favorito, Giuseppe Galano, Sara Giammusso, Antonio Muci and Matteo Nardelli*

n. 51 Environmental data and scores: lost in translation, *by Enrico Bernardini, Marco Fanari, Enrico Foscolo and Francesco Ruggiero*

n. 52 How important are ESG factors for banks' cost of debt? An empirical investigation, *by Stefano Nobili, Mattia Persico and Rosario Romeo*