Review of Quantitative Finance and Accounting (2024) 63:925–954
https://doi.org/10.1007/s11156-024-01278-0

ORIGINAL RESEARCH

Can we trust machine learning to predict the credit risk of small businesses?

Alessandro Bitetto[1] · **Paola Cerchiello**[1] · **Stefano Filomeni**[2] · **Alessandra Tanda**[1] · **Barbara Tarantino**[1]

Accepted: 28 March 2024 / Published online: 6 June 2024

Abstract
With the emergence of Fintech lending, small firms can benefit from new channels of financing. In this setting, the creditworthiness and the decision to extend credit are often based on standardized and advanced machine-learning techniques that employ limited information. This paper investigates the ability of machine learning to correctly predict credit risk ratings for small firms. By employing a unique proprietary dataset on invoice lending activities, this paper shows that machine learning techniques overperform traditional techniques, such as probit, when the set of information available to lenders is limited. This paper contributes to the understanding of the reliability of advanced credit scoring techniques in the lending process to small businesses, making it a special interesting case for the Fintech environment.

Keywords Small businesses · Credit rating · Credit risk · Invoice lending · Machine learning · Fintech

JEL classification C52 · C53 · D82 · D83 · G21 · G22

1 Introduction

Small businesses have always struggled to obtain funding through traditional channels because of their limited size and high information asymmetries (Sharpe 1990; Ivashina 2009). This acts as an obstacle to their potential growth and development in the marketplace (Berger and Udell 2006).

Among the financing channels, bank lending has always received great attention by the extant literature (Agostino et al. 2012; Canales and Nanda 2012; Grunert and Norden 2012; Beck 2013). Indeed, especially in the past, banks were able to overcome information asymmetries through "relationship lending" that allowed them to incorporate borrower-specific soft information in the lending process, thus enabling informationally opaque small businesses

✉ Stefano Filomeni
stefano.filomeni@essex.ac.uk

[1] University of Pavia, Department of Economics and Management, Pavia, Italy

[2] University of Essex, Essex Business School, Finance Group, Colchester, UK

to be financed (Filomeni et al. 2021). After the regulatory evolution of the Basel Accords, the employment of soft information has become more and more difficult due to limitations to hardening soft information in banks' internal ratings, thus causing a shift in the preferences of banks towards alternative ways to assess corporate creditworthiness that left small businesses lending demand somehow unmet (Berger 2006; Filomeni et al., 2021; 2020).

In this context, policymakers and small businesses have welcomed the birth and diffusion of Fintech (financial technology) and the application of advanced methodologies in the field of banking and finance. The Fintech revolution has brought new ways to interact with small businesses that can finally have a timely response to their financing needs by new and old lenders (Tanda and Schena 2019; Gong and Ribiere 2021).

Fintech companies developed especially in the segment of invoice lending (sometimes referred to as "trade credit") (Dorfleitner et al. 2017). Invoice lending enables firms to obtain new resources via the presentation of invoice receivables or other credit instruments to be discounted (Soufani 2002). Firms that issue invoices (creditors to their customers) can therefore apply for the "anticipation" of the discounted amount exhibited on the invoice (or another credit instrument), essentially employing the unpaid customer invoices as collateral to obtain immediate cash from a lender. It is a form of short-term borrowing that provides businesses with quick access to working capital. It is a widespread form of financing that enables businesses to access funds that are tied up in outstanding invoices, providing them with immediate liquidity to meet their operational needs, such as covering expenses, investing in growth opportunities, or managing cash flow gaps. Invoice lending therefore allows companies to bridge the time mismatch between invoicing their customers and receiving payment from them.

Fintech platforms allow especially small firms to use this type of supply chain financing instrument via different business models (International Monetary Fund 2017, Schena et al. 2018). It is possible that these platforms provide a model of "direct" financing selection by individual customers (in this case, the platform may select applications according to scoring techniques or acceptance criteria). Alternatively, as described by the International Monetary Fund (2017), the platform may apply a "diffused" model, in which case it will group funding applications into homogeneous classes, thus having a greater effect on the creditworthiness assessment of the borrower and, consequently, on the funding choices and risk borne by the lenders. Depending on the business model adopted, Fintech lenders can collect resources to finance invoice lending from the crowd of investors or from specialized financial intermediaries, including banks or other traditional lenders.

New Fintech platforms have become especially successful for their ability to make timely decisions on lending requests and their speed at responding to customers' needs. This is made possible by their lean structures and also their ability to employ advanced techniques to overcome or mitigate the information asymmetries that generally affect financial transactions. For a few years now, also traditional lenders have started to adopt these advanced methodologies, which include artificial intelligence (AI) and machine learning (ML) algorithms. These innovations however have not come without risks (Gomber et al. 2018; Ozili 2018; Thakor 2020).

Indeed, despite the diffusion of ML, not all ML methods are deemed to be reliable or applicable by policymakers and regulators. In the financial sector, many have underlined the need for explainable algorithms especially when dealing with savings and financing decisions (The Royal Society 2019, Hadji-Misheva and Osterrieder 2023) and more transparency in relation to Fintech lending has been advocated. Within this framework, this paper aims at evaluating if innovative machine learning techniques, based on artificial intelligence, can contribute to an accurate credit risk evaluation for small businesses. To this end, we employ and compare alternative machine learning algorithms to predict the credit rating in

the framework of invoice lending by exploiting a unique proprietary dataset of securitized invoices to be payable by Italian small- and medium-sized businesses.

We select this environment to frame our empirical analyses as it is currently the most successful type of lending issued by Fintech companies, especially in Italy (Financial Stability Board 2017; Dorfleitner et al. 2017; Zhang et al. 2016).

This paper contributes to the literature on small business lending in three different ways. First, we test the capability of ML techniques to predict the credit risk of small firms in the context of digital financial markets (Financial Stability Board 2017; Dorfleitner et al. 2017; Bitetto and Cerchiello 2023). Despite the AI nature of the methodology, we are also able to provide meaningful insights into the drivers of our results, by employing Shapley values to open the "black box". Shapley values are a concept taken from cooperative game theory used to fairly distribute the total payoff of a collaborative effort among the contributing participants. Therefore they can be applied to black box models to assess the average marginal contribution of each player, i.e., the predictor variable, to the payoff, i.e., the predicted output. This approach adds value to both the borrowers that can better understand how they can improve their fundamentals to access more and cheaper credit, and policymakers that strongly support the need for explainable artificial intelligence algorithms (The Royal Society 2019).

Second, to show the outcomes of ML techniques, we report the perspective of the Fintech lenders that may have limited access to information and are therefore able to provide the interested audience (regulators, policymakers, financial institutions) with a crystal clear representation of the plausible results you can get according to the model/available data within the Fintech environment.

Third, to evaluate the validity of ML in the face of traditional models, we compare the ML and probit methodologies using a set of different models, according to the pool of information available: (i) only financial statement information (FS); (ii) only information related to invoice payment behavior (INV); (iii) both sets (FS + INV). These three pools of information are investigated in a case where the lender has only the most recent information (contemporaneous information) and in a case where the lender has also access to the historical information.

Our empirical findings show that ML performs better with limited contemporaneous information, while ML and probit perform similarly when lenders have access to historical data for FS items. Moreover, probit outperforms ML only when complete historical information on financial statement items and behavior on payments is available. Finally, we prove that the most relevant variables for FS information set are Turnover, together with Current liabilities, Working capital, and EBIT. With regards to the INV information set, i.e., the variables referring to the firm's invoice payment activity, we provide evidence that Delinquency and Outstanding contribute the most to the model prediction capability. Our results hence call for the inclusion of ML methodologies in the invoice lending environment, especially for small firms' credit risk evaluation, wherein Fintech platforms are likely not to have access to historical or behavioral information.

The remainder of the paper is organized as follows: Section 2 presents a review of the literature; Section 3 describes the data and the empirical methodology; Section 4 discusses the results of our empirical analysis; Section 5 concludes the paper.

2 Literature review

Small businesses are characterized by informational opaqueness and greater perceived risk. The availability of information affects their default prediction and credit rating, a long debated topic in the literature (Ciampi et al. 2021, among others). This represents an issue for financial

intermediaries that have to estimate firms' creditworthiness based on hard information that enters a credit rating (or credit scoring) model that returns a credit rating (or score). On the basis of the assigned credit rating, the whole lending process develops determining small business availability of funds. This process is particularly sensitive, especially when it is difficult to estimate the borrowers' creditworthiness due to informational opaqueness and during crisis periods. Indeed, as discussed by Duarte et al. (2018) and Ciampi et al. (2021), financial and economic crises have the effect to reduce the amount of funds available, especially to SMEs , as the latter are "physiologically more opaque" than other companies (Ciampi et al. 2021).

Especially in the past, information asymmetries were mitigated by banks thanks to the continued relationship with their borrowers that allowed banks to collect "soft information". This information advantage constituted an important competitive advantage of incumbent lenders over new banks (Berger and Udell 1995; Dell'Ariccia 2001; Claessens et al. 2005; OECD 2020; Filomeni et al. 2023a). Over time, banks' ability to rely on soft information has become thinner because of changing regulations and more complex organizational frictions that prevent the successful hardening of soft information (Stein 2002; Liberti and Petersen 2018; Filomeni et al. 2021; 2020; Bitetto et al. 2023). These changes left a share of loan demand by small businesses unmet. This paved the way for the search and the development of alternative sources of financing for small businesses.

The phenomenon of digitalization of financial markets, commonly known as Fintech, has brought new opportunities for small businesses. New providers of funds and new platforms (e.g., P2P platforms and crowdfunding platforms) allow small businesses to access funds, often through a very lean and timely process (Tanda and Schena 2019) and increase the options and the credit available to SMEs. Both Beaumont et al. (2022) and Gopal and Schnabl (2022) find that Fintech platforms provided an additional lending source to small businesses during the crisis. Abbasi et al. (2021) also show, over a more recent period, that Fintech lenders increase credit availability to SMEs. These platforms often rely on methodologies to cope with limited information based on advanced methodologies, including artificial intelligence (AI) and machine learning (ML) algorithms.

Both the industry and the academic literature have studied methodologies to estimate corporate credit risk employing AI and ML methods that, however, are "explainable", i.e., that allow borrowers, lenders, and regulators to understand the drivers of the results obtained through AI evaluation methods for creditworthiness assessment. For instance,' Breeden (2021) has recently provided an overview of machine learning in credit risk and underlined that explainability is essential to have regulatory-compliant machine learning methods in the area of financial services to couple the flexibility and accuracy of non-linear models with the need to protect the stakeholders in financial contracts. Medianovskyi et al. (2022) argue that explainable AI tools are able to limit information asymmetries between SMEs and potential lenders.

Beside the necessity to build and apply transparent methodologies, the literature has also discussed the capability of Fintech companies to correctly evaluate the creditworthiness of small businesses. Fintech companies might lack the necessary skills and access to data which are required to estimate firms' creditworthiness (Kowalewski and Pisany 2022). This represents a serious policy concern, as the funds channelled from investors to companies should be directed to the best borrowers and, hence, investors should be able to differentiate between good and bad borrowers. An efficient allocation of resources is a primary goal of regulators and policymakers and one key function of financial intermediaries and the financial system. The ability of machine learning to correctly evaluate corporate creditworthiness is, hence, of primary importance (Byanjankar et al. 2015). For instance, Nguyen et al. (2023) show that machine learning algorithms coupled with methods to explain results, such as Shapley Additive Explanations (SHAP) can be extremely useful to predict

the creditworthiness of borrowers. Altman et al. (2023) developed an alternative method to the well-known Altman Z-score (Altman 1968; Altman et al. 2017) that employs machine learning to estimate SMEs' probability of default, computed not only on the basis of financial ratios but also on other qualitative measures, including, for instance, employees' and managers' characteristics. Additionally, these methodologies allow the borrower and the lender to understand the drivers of the outcome, thus rendering explainable artificial intelligence methods preferable to other approaches.

3 Data and methodology

3.1 Data

The initial raw dataset contained a panel of invoice data on 534 Italian small- and mid-sized businesses (henceforth "small firms") for the period that spans from the first quarter of 2015 to the second quarter of 2017.[1]

Data comprises information on invoice transactions (invoice-related data, henceforth INV) collected from a large European bank specializing in revolving trade receivables' securitization programs. Credit ratings are indeed assigned to small- and mid-sized businesses in the framework of a securitization program initiated by the bank involving revolving trade receivables in favor of some of its corporate clients. Through credit ratings, the bank can indeed assess the credit risk of the acquired portfolio of securitized trade receivables originated by its corporate clients engaged in invoice financing that improve their corporate cash flow by getting immediate access to the cash from the invoice they issue on completion of the job, instead of waiting the normal 30 to 60 + days for payment. Securitization is the process in which trade receivables originated by the bank's corporate clients are pooled so that they can be repackaged into interest-bearing securities, typically purchased by the bank itself to which the payments from the securitized trade receivables are passed through (Filomeni, 2024). These corporate clients are those originating the trade receivables to be securitized (i.e., the originators), whereas their clients are identified as the "final debtors" in the securitization process as they are referred to as the obligors who owe the originators payments on the underlying trade receivables and are, therefore, ultimately responsible for the performance of the securities as part of the securitization transaction. Therefore, the data used in this paper can be divided into invoice-related (INV) and financial statement data (FS). Data related to invoice lending were gathered from a prominent European institution involved in revolving trade receivables' securitization programs. Accounting information was obtained from the Orbis database (Bureau Van Dijk, a Moody's analytics company), by associating the VAT code for each given "final debtor".

Data are consistently collected across a sample of invoice transactions reflecting small and mid-corporate debtors managed by a major European bank specializing in the revolving trade receivables' securitization programmes over the period Q1 2015-Q2 2017. Our European bank belongs to a large European banking group and is representative of the general population of banks in the Eurozone. At the time of the data collection, the group had total assets of around 650 billion euros, it is publicly listed, and has a market capitalization reaching 50 billion euros. Its subsidiaries are located in twelve European countries, mainly located in the Southern, Central and Eastern parts of the continent. The Italian country territory represents one of the major markets where the bank operates. The bank lending activity in this country can be considered as representative

[1] We acknowledge that the empirical analysis of this study covers a limited time period; however, the literature has extensively used similar granular proprietary data with a limited time period (e.g., Liberti and Mian (2009), Filomeni et al. (2021; 2020; 2023a)).

of the overall banking industry, covering a market share of about 15% in the loan and deposit markets, while the group is present with 14 affiliated banks and about 4500 branches spread out over the country under analysis. We therefore contend that our empirical results can be extended to other banks as our data-providing bank is highly representative of the banking industry, thus ruling out possible bank-specific idiosyncratic issues that may characterize the single banking organization.

Firms subject to invoice trading (i.e., buyers of goods or services in the trade) are privately held companies, firstly identified through their VAT code. In the invoice lending environment, the relevant credit rating to be estimated by our invoice purchasing bank is the final debtor rating, i.e., the rating of the buyer of the goods that shall pay the invoice presented by the seller. To evaluate the economic and financial profile of the final debtor, we collect accounting information on key financial statement items of debtor firms from the Orbis database (developed by Bureau Van Dijk), by matching the VAT code for each given final debtor.

Our final dataset, therefore, includes two types of data: variables describing the invoices and information on accounting data of the final debtors.

In detail, 6 numerical variables referred to as invoice lending (hereinafter "INV" data) were provided by the financial intermediary with quarterly frequency and 28 (25 numerical and 3 categorical) financial statement variables (hereinafter "FS" data) were collected by Orbis platform with annual frequency.[2] Annual values are repeated over all quarters of each year.

Variables refer to corporate credit quality (Rating), financial statement key information (e.g., Current liabilities, EBIT, Fixed assets), payment behavior (e.g., Delinquency and further transformations), firm characteristics, including area and industry dummies.

Additionally, Nace Rev. 2 is utilized to categorize the primary sector (NACE) and primary division (Industry) of the firms. Geolocalization variables have been built through Google Maps API and have been linked to each firm present in the dataset to control for unobserved heterogeneity associated with the given firm's industry and location. Table 1 presents the definition of the variables employed in the empirical analysis along with their descriptive statistics.

The initial dataset was cleaned and checked for outliers, redundant data, and missing values. To improve the manageability of data, the variables have been normalized with respect to a different set of features, according to the nature of the specific data (invoice-related or financial statement data) in order to obtain a normalized set of predictors between 0 and 1. However, some extreme values have been deliberately left in the dataset to reflect the extreme characteristics of some firms with respect to the normalized range and to avoid having a dataset with too few observations.

Moreover, outliers were eliminated using the inter-quantile range (α–quantile and $(1-\alpha)$–quantile). However, to preserve values of variables with small variance, instances where the distance between the maximum and minimum values was less than a specified tolerance were retained, and no outliers were removed. Categorical variables were transformed into dummy variables, excluding the $n-th$ level in order to avoid multicollinearity.

The correlation among the FS and INV variables was examined, resulting in the removal of 8 variables with a Variance Inflation Factor (VIF) exceeding the value of 5. The presentation of the distribution of dummy and categorical variables for each credit rating is provided in Table 2.

The final dataset consists of 464 firms and 21 variables, of which 6 INV and 15 FS, treated according to an unbalanced panel data structure resulting in 3,009 rows.

[2] Infra-annual data in Orbis for SMEs or smaller companies are very difficult to retrieve.

Table 1 List of final variables

Variable	Description	Avg	Std. dev	50th percentile	Minimum	Maximum	Removed due to VIF
Rating variables							
Rating	Rating score, 2 means high credit worthiness	5.11	1.24	5.00	2.00	9.00	
Financial Statement variables (FS)							
Purchase	Accounting of Cash and Credit purchases	1.49	0.96	1.31	0.02	6.48	
Current liabilities	Company's debts or obligations that are due to be paid to creditors within one year	0.55	0.20	0.55	0.04	2.03	
Current ratio	Ratio of a firm's current assets to its current liabilities	0.01	0.01	0.01	0.00	0.19	x
EBIT	Company's net income before income tax expenses and interest expenses are deducted	0.05	0.09	0.04	-1.44	0.69	
Fixed assets	Long-term tangible piece of property or equipment that a firm owns and uses in its operations	0.33	0.22	0.30	0.00	0.98	x
Liquidity	Company's ability to pay off current debt obligations without raising external capital	0.01	0.01	0.01	0.00	0.16	
Turnover	Annual sales volume net of all discounts and sales taxes in logarithmic scale (base 10)	4.53	0.83	4.44	2.85	6.94	
LT Debt	Debt with maturities greater than 12 months	0.09	0.10	0.05	0.00	0.52	
Asset Turnover	Sales revenue divided by capital employed	0.08	0.18	0.04	0.00	3.53	x
Profit Margin	Percentage of sales turned into profits	0.02	0.06	0.02	-0.73	0.56	
Profit per employee	Net Income for the past twelve months (LTM) divided by the current number of Full-Time Equivalent employees	0.00	0.05	0.00	-0.02	1.00	
ROA	Net income divided by total assets	0.03	0.09	0.02	-0.35	1.92	
ROCE	Company's earnings before interests and taxes (EBIT) divided by capital employed	0.08	0.22	0.07	-7.31	0.85	x
ROE	Fiscal year net income divided by total equity	0.08	0.64	0.08	-13.72	9.73	
Solvency	Firm's capacity to meet its long-term financial commitments	0.28	0.18	0.25	-0.79	0.93	
Tangibles	Assets that have a physical value	0.25	0.19	0.21	0.00	0.98	
Working Capital	Difference between a company's current assets and current liabilities	0.14	0.24	0.12	-1.72	1.07	

Table 1 (continued)

Variable	Description	Avg	Std. dev	50th percentile	Minimum	Maximum	Removed due to VIF
Variables related to invoice financing (INV)							
Delinquency	Dummy variable equal to 1 if the firm misses a scheduled payment on an invoice and 0 otherwise	0.02	0.10	0.00	0.00	1.00	
Collections	Amount of invoices currently paid to the purchasing bank	2.71	90.70	0.77	0.00	5520.70	
Outstanding	Amount of invoices stemming from the securitization transactions in which the bank's client is involved, expressing its economic exposure in logarithmic scale (base 10)	4.16	1.95	4.68	0.00	7.18	
New Receivables	Monetary amount of receivables sold to the bank from a given bank's client at the current invoices' transfer	0.21	0.24	0.16	0.00	1.00	
Outstanding_Invoices	Amount of invoices stemming from the securitization transactions in which the bank's client is involved divided by the total number of invoices	0.13	0.34	0.00	0.00	1.00	x
Outstanding_Portfolio	Amount of securitization transactions in which the borrowing firm is involved divided by total number of portfolios	0.15	0.35	0.00	0.00	1.00	x
Delinquency Severe	Dummy variable equal to 1 if the delinquency (payments overdue) amount is larger or equal than + 2 standard deviations from the average of all clients	0.04	0.21	0.00	0.00	1.00	
Delinquency 90	Dummy variable equal to 1 if Scaduto90 (i.e., payments overdue by more than 90 days evaluated on average by ID) is larger than 0 and 0 otherwise	0.29	0.45	0.00	0.00	1.00	
Liquidity Tension	Dummy variable equal to 1 if Collectionperioddays (i.e., number of days it takes to turn accounts receivables into cash) is larger than Creditperioddays (i.e., number of days that a customer is allowed to wait before paying an invoice) and 0 otherwise	0.43	0.49	0.00	0.00	1.00	x
Other controls							
NACE	Statistical Classification of Economic Activities in the European Community						
Industry	Industrial classification reflecting the firm's main division within the main section of NACE						x
Region	Geographical macro-areas						

Table 2 Dummy and categorical variables distribution by each credit rating

Variable		Rating							
		2	3	4	5	6	7	8	9
NACE									
Manufacturing		20%	23%	28%	34%	35%	31%	18%	0%
Wholesale and retail trade; repair of motor vehicles and motorcycles		57%	61%	54%	53%	54%	59%	82%	100%
Accommodation and food service activities		22%	8%	15%	10%	7%	8%	0%	0%
Agriculture, forestry, and fishing		2%	7%	1%	1%	2%	1%	0%	0%
Other		1%	6%	3%	2%	2%	3%	0%	0%
REGION									
North East		24%	35%	42%	38%	26%	20%	8%	0%
North West		61%	35%	28%	23%	26%	25%	31%	0%
Center		6%	20%	16%	16%	19%	29%	12%	25%
South and Islands		9%	10%	15%	22%	30%	26%	49%	75%
DUMMY VARIABLES									
Delinquency Severe	0	74%	90%	92%	98%	100%	100%	100%	100%
	1	26%	10%	8%	2%	0%	0%	0%	0%
Delinquency 90	0	47%	29%	60%	74%	82%	82%	43%	25%
	1	53%	41%	40%	26%	18%	18%	57%	50%
Liquidity Tension	0	76%	69%	68%	54%	50%	56%	59%	100%
	1	24%	31%	32%	46%	50%	44%	41%	0%

3.1.1 Our key dependent variable

Within the invoices-related data, a significant metric representing a borrower's credit quality is represented by the credit rating. The credit rating, assigned by the financial intermediary to each final debtor, assesses borrowers' financial well-being and creditworthiness. Specifically, it predicts the likelihood of corporate default by analyzing both proprietary soft information (such as client details and special investigations) and private/public hard information (including partnerships, registered payment defaults, credit reference agency data, accounting information, payment performance data, and risk network information).[3] Consequently, the credit rating offers an objective and measurable way to evaluate firms' credit risk.

In the context of our study, particular relevance is attributed to mid- and small-sized businesses' credit ratings. The latter are assigned to small businesses by a single financial intermediary that uses them to assess their corporate credit risk. Small businesses' credit ratings make them suitable to the purpose of our study. Indeed, attributed credit ratings are based on both hard and soft information. On the one hand, the former is based on private and publicly available quantifiable information (i.e., partnerships, registered payment

[3] Credit scores are not fixed and can be influenced by various factors. There are several strategies to raise low scores and potentially reduce premiums. One effective approach is to enhance one's credit rating by ensuring timely payment of bills and reducing overall debt. Additionally, within a securitization involving insurance companies, limiting the number of filed insurance claims within a specific period can contribute to improving an insurance score.

defaults, credit reference agencies, accounting data, payment performance data, network of risk information). On the other hand, the latter is based on relationship-intensive soft information collected directly and indirectly through repeated bank-firm interactions. In the context of our study, the credit rating (the target feature) is a factor variable with eight categories ranging from a minimum value of 2 to a maximum value of 9, where the 9th rating class represents the riskiest one characterizing the least creditworthy borrowers. The different notches of the rating scale allow the financial intermediary to distinguish between high-risk and low-risk clients in their lending activity. Therefore, credit ratings provide an objective and quantifiable means by which a company's degree of credit risk can be assessed.

Credit rating evolution over time is shown in Fig. 1, highlighting an overall persistent behavior for all classes of risk.[4]

3.2 Methodology

To investigate the ability of machine learning (ML) to provide reliable credit ratings' predictions, we rely on a proprietary dataset comprising firms involved in invoice trading for which we have information on the assigned credit rating, key financial statement data, and information on their behavior in terms of payment of invoice, delays, overdue, etc.

Given that our focus is primarily to study the viability and accuracy of ML techniques in the Fintech environment, we suppose that invoice lenders can have access to a diversified pool of information, which depends on the business model adopted by the Fintech lender. The sets of information are built as follows: set (i) comprises only financial statement information (FS); set (ii) comprises only contemporaneous information on the pool of invoices and payment behavior (INV); set (iii) comprises both financial statement and invoices-related information (FS + INV). Additionally, we include a time factor, differentiating between lenders that can access only contemporaneous information (time t) and lenders that can access historical information (time s, with $s < t$). The dependent variable is represented by an ordinal variable and therefore we select a random forest for our machine learning approach. Results are then compared with those derived from an ordered probit model. We start by describing our implemented random forest approach.

Random forest (RF) applications represent a non-parametric machine learning method that has now been widely employed in many fields of academic research (Breiman 2001). RF models primarily rely on an ensemble of decision trees, which is recognized as a cutting-edge machine learning approach for prediction and classification tasks (Biau and Scornet 2016). We apply RF to a dependent variable, i.e., the credit rating (y_{it}) that is modelled as a prediction of a set of independent variables X_{it}. In coherence with the ordinal nature of our dependent variable y, we employ the classification version of RF model.

The first approaches dealing with longitudinal and clustered data involved tree-based methods (Segal 1992; Sela and Simonoff 2012; Hajjem et al. 2014) and are based on the idea of iterating between fixed and random parts and estimating the parameters via the Expectation Maximization (EM) algorithm. All these approaches represent a semi-parametric fixed effects model in which the non-parametric part is evaluated through RF. For the sake of full comparison and robustness check, we also employ a modified random forest algorithm that includes summary transformations of the past values for every input variable, named Historical Random Forest (HRF). Therefore, HRF is suitable for the analysis of longitudinal data. In the *R* environment, we employ the package htree (Sexton 2018).

[4] This shows no particular shocks in the overall economic system during the time period considered. Our results are therefore valid during normal times.

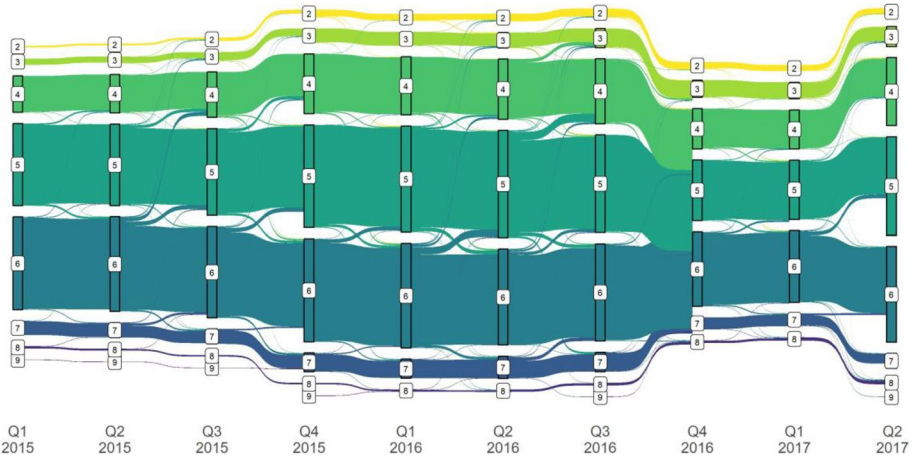

Fig. 1 Rating evolution over time

The relationship between the credit rating (y) and the set of regressors is also tested by implementing an ordered probit model according to Eq. 1.[5]

$$y_{it} = \mathrm{X}_{it}\beta + \alpha_i + \varepsilon_{it} \tag{1}$$

where i indicates the firm and goes from 1 to N; t indicates the quarter and ranges from 1 to T; consequently, y_{it} is our dependent variable, i.e., the value of the credit rating assigned to each firm i in each quarter t ranging from 2 (best creditworthiness) to 9 (lowest creditworthiness); X_{it} is a vector of dimension $1 \times k$ ($k=21$) containing our explanatory variables at time t; β is a vector that contains the parameters to be estimated; α_i is a constant, i.e., firm-specific and time-invariant component; ε_{it} is interpreted as the disturbance term, which we assume to be normally distributed. According to the business model adopted by the invoice lender, the latter could also access historical data. Additionally, given that rating changes are subject to serial correlation and are not updated timely with new information becoming available (Gonzalez et al. 2004; Odders-White and Ready 2006), we include the lagged value of the rating in Eq. 1, that results now in Eq. 2. The dependent variable is modelled as a first-order Markov process, following Contoyannis et al. (2004), Wooldridge (2005) and Greene and Hemsher (2008).

$$y_{it} = \mathrm{X}_{it}\beta + y_{i(t-1)}\gamma + y_{i0}\delta + \alpha_i + \varepsilon_{it} \tag{2}$$

where $y_{i(t-1)}$ indicates the lagged value of the credit rating for firm i, γ represents the parameters associated to the rating in the previous time period, y_{i0} is the first available firm's rating at time $t=0$.

To select the best combination of variables in the probit environment, we employ a set of well-known evaluation metrics. After looking into the confusion matrix, we select the F_1-score that shows the best value, choosing between (i), (ii), and (iii), namely $F_{1\text{cross-entropy}}$ with $\gamma=4$.

(i) $$F_{1_{ratio}} = F_{1_{test}} + \frac{F_{1_{test}}}{\Delta F_{1_{train-test}}}$$

(ii) $$F_{1_{harmonic}} = \frac{2}{\frac{1}{F_{1_{test}}} + \frac{1}{\Delta F_{1_{train-test}}}}$$

[5] To perform our analysis with the probit models we employ the R package oglmx by Carroll (2018).

(iii) $F_{1_{cross-entropy}} = -F_{1_{test}}^{\gamma} \log\left(1 - F_{1_{test}}\right) - \left(1 - \Delta F_{1_{train-test}}\right)^{\gamma} \log\left(\Delta F_{1_{train-test}}\right), \gamma \geq 1$

To validate the performance of the model and the split into train and test samples, we employ a variable-length rolling-window temporal approach. Specifically, considering that the maximum number of quarters available in our sample period is 10 and the total number of quarters for each firm varies, a test set comprising the two most recent quarters has been selected, while the remaining quarters constitute the training set. Since each firm has a minimum of 7 available quarters and we have set a minimum of 10 observations in each training set, a total of 4 folds are used in the cross-validation process.

3.3 Variables importance assessment

We address the assessment of the predictive power of the variables by employing two different techniques. The first is the Permutation Feature Importance (PFI), which involves measuring the change in the model's prediction error when the feature's values are shuffled. If permuting the values leads to an increase in the prediction error, the feature is deemed relevant for the model's prediction. Conversely, if the model's error remains unchanged, the feature's contribution is considered unimportant.

Following the proposal by Fisher et al. (2018), the algorithm for a generic model f can be defined as in Algorithm 1.

Algorithm 1 Permutation Feature Importance (PFI).

Input: Trained model f, feature matrix X, target vector y, performance metric $P(y, f)$

Estimate the original model performance $P_{\text{orig}} = f(y, X)$;

foreach *feature* $j = 1, \ldots, p$ **do**

  Generate feature matrix X_{perm} by permuting feature j in the data X;

  Estimate $P_{\text{perm}} = f(y, X_{\text{perm}})$ based on the predictions of the permuted data;

  Evaluate $\text{PFI}_j = P_{\text{perm}} / P_{\text{orig}}$. Alternatively, the difference can be used:

  $\text{PFI}_j = P_{\text{perm}} - P_{\text{orig}}$;

  return PFI_j;

end

Sort features by descending PFI

The second method we employ to rank variables according to their importance in the prediction of the credit rating is the SHAP values method. Shapley values quantify the individual contributions of each feature to the prediction of a specific data point . In this context, the feature values, such as those of instance x, can be likened to players in a game where the prediction serves as the payout. More precisely, the Shapley value Φ_j of a feature value x_j is determined using a value function *val* for actors in S, and it represents the weighted contribution of the feature value to the prediction across all potential coalitions (as shown in Eq. 3) (Shapley 1953; Bussmann et al. 2021).

$$\Phi_j(val) = \sum_{S \subseteq \{x_1,\ldots,x_p\}\{x_j\}} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup \{x_j\}) - val(S)) \tag{3}$$

where S denotes a subset of features (or independent variables), x represents the feature values of the variable of interest, p the number of features, and $val_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in S (Eq. 4):

$$val_x(S) = \int \hat{f}(x_1, \ldots, x_p) d\mathbb{P}_{x \notin S} - E_X \hat{f}(X) \tag{4}$$

To overcome the computational challenge deriving from estimating the Shapley values for many features we apply the Strumbelj and Kononen's (2014) Monte-Carlo sampling process (Eq. 5).

$$\hat{\Phi}_j = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}\left(x_{+j}^m\right) - \hat{f}\left(x_{-j}^m\right)\right) \tag{5}$$

where $\hat{f}\left(x_{+j}^m\right)$ represents the prediction for the instance of interest x but with a random permutation of features (taken from a random data point z) with the exclusion of the j-th feature; $\left(x_{-j}^m\right)$ is a vector identical to $\left(x_{+j}^m\right)$, except for the fact that the value for feature j is randomized, as well as it is the one from the sampled z. The algorithm for a generic model f can be defined as in Algorithm 2, that shall be repeated for every feature.

Algorithm 2 Shapley value.

Output: Shapley value for the value of the j-th feature

Input : Number of iterations M, instance of interest x, feature index j, data matrix X, and machine learning model f

foreach $m = 1,\ldots,M$ **do**

Draw random instance z from data matrix X;

Choose a random permutation o of the feature values;

Order instance x: $x_O = (x_{(1)},\ldots,x_{(j)},\ldots,x_{(p)})$;

Order instance z: $z_O = (z_{(1)},\ldots,z_{(j)},\ldots,z_{(p)})$;

Construct two new instances:

- With feature j: $x_{+j} = (x_{(1)},\ldots,x_{(j-1)},x_{(j)},z_{(j+1)},\ldots,z_{(p)})$
- Without feature j: $x_{-j} = (x_{(1)},\ldots,x_{(j-1)},z_{(j)},z_{(j+1)},\ldots,z_{(p)})$

Compute marginal contribution: $\Phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$;

return Φ_j^m;

end

Compute Shapley value as the average: $\Phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \Phi_j^m$

With respect to other AI methods, Shapley values offer an advantage as they enable the assessment of each explanatory variable's contribution to individual point predictions of a machine learning model, irrespective of the specific model being used (Lundberg and Lee 2017). More clearly, Shapley-based explainable AI models combine the flexibility to be applied across various models (being model-agnostic) with the ability to provide personalized explanations for each individual prediction.

3.4 Statistical assessment of differences

Following the assessment of multiple learned classifiers and the subsequent classification of new samples with unknown class labels, it becomes imperative to conduct a statistical comparison of classifiers. This comparison is essential to evaluate the statistical differences between the results obtained by different algorithms across various instances of problems, datasets, and more. The typical analytical sequence begins with the application of a test that simultaneously compares all the considered algorithms to test for any instances where an algorithm behaves differently. If the null hypothesis is rejected, signifying globally significant differences, the subsequent step involves analyzing which pairwise combinations exhibit differences through the implementation of post-hoc tests.

Initially, the classical non-parametric Friedman test (Friedman 1937) was applied. In cases where observations do not adhere to measurement requirements, and to sidestep assumptions about the underlying populations, non-parametric statistical tests become appropriate, with Friedman's test being a notable choice. This test serves as a non-parametric alternative to the parametric twoway analysis of variance, aiming at identifying differences in treatments across multiple test attempts. The computational process entails ranking each row collectively, arranging the values of the row in decreasing order, and calculating the average rank for each column. The formula to compare two columns is the following:

$$z = \frac{\left(R_i - R_j\right)}{\sqrt{\frac{k(k+1)}{6N}}} \tag{6}$$

where R_i is the average rank obtained from the Friedman test for column i, k represents the number of columns, and N is the number of block sets, both used for comparison purposes. The fundamental concept is to compare the accuracy of various classifiers using different datasets. Consequently, the columns represent the classifiers and the rows correspond to the datasets.

Then, the corresponding post-hoc tests for Friedman have been implemented, correcting p-values for multiple testing (i.e., Bergmann and Hommel's correction procedure). The latter applies a correction based on a list of possible hypothesis testing and amplifies the test power by considering only exhaustive sets of hypotheses (i.e., hypotheses that can be simultaneously true).

4 Empirical analysis

We first run some classification performance tests on our subset of information pools. We employ the macro-averaged F_1-score to assess the models' predictions.

To this end, we used a set of evaluation metrics (shown in Appendix 1) in order to obtain an optimal combination of hyper parameters and variables. We employ the F_1

cross-entropy metric and the well-known Akaike Information Criterion or AIC.[6] The former has been maximized to avoid over fitting and the latter has been minimized to ensure the stability of predictors (see Table 9 for an overview of the hyper parameters set and the set of predictors selected with reference to the invoice set of predictors INV).

Table 3 reports the classification performance with regard to both the training and test samples. Results in the table are differentiated according to the pool of information available to the invoice lender, namely models that consider: (i) FS only, (ii) INV only, and (iii) FS + INV, estimated with (a) only contemporaneous information at time t and (b) with the historical information. Results show that random forest (RF) has a quite good performance, especially when we consider the pool of information (iii), i.e., FS + INV. On the contrary, the probit model (PB) has a very low performance when historical information is not available (scenario a in Table 3) across all sets of information (i-iii). These findings hold immense significance as they highlight the limited utility of classic parametric models in situations where information is limited. In such cases, more advanced techniques like ML play a crucial role in leveraging the limited available data by effectively capturing nonlinear relationships. These sophisticated approaches are essential for maximizing the potential of the limited information set.

In the alternative scenario (b) in Table 3, when the historical informative set is available, ML and PB perform similarly. Both approaches result to be competitive and they do not significantly differ. Furthermore, we observe a notable enhancement in overall performance with the inclusion of historical data, as evidenced by a significant increase in the F_1-score.

If we focus on the contribution of each group of variables, in Table 3 we notice that FS and INV are comparable with a slightly better performance of the former. Based on the results regarding variable importance and best subset selection, which aimed at reducing the high-dimensional feature space and obtain an optimal group of features, a final model was implemented by combining both FS and INV sets of variables. Before fitting the models, a preliminary analysis was conducted to check for correlation and collinearity. The results revealed a significant correlation between Outstanding and Turnover. Consequently, Outstanding was removed since a measure of a firm's financial exposure had already been selected with Delinquency, and having a metric for firm's size proxied by generated revenues appeared useful in determining the credit rating. Although the regional and industrial classification variables show a significant effect on the target for one specific category and the impact could not be confirmed in terms of importance, both regional and industrial classification variables were retained to provide insights into the economic framework. In summary, the following variables were selected for the final set: *Turnover, Solvency, Working Capital, LT Debt, Current liabilities, Liquidity, Collections, New Receivables, Delinquency, NACE, Region*.

4.1 Model explanation

Besides a strict performance evaluation, a more comprehensive one is crucial in terms of implications induced by the considered variables. Since parametric (PB) and non-parametric (RF/HRF) models are not directly comparable in terms of contribution and relevance of the employed variables, we set a common ground of comparison through Shapley values and PFI. In other words, we need to render the ML approaches explainable

[6] The AIC is designed to strike a balance between model accuracy and parsimony, favouring models that fit the data well while using fewer parameters (Akaike 1974).

Table 3 Macro-averaged F_1-score on training and test samples for all set of predictors

			F_1-score		
Model	Version	Sample	FS (i)	INV (ii)	FS + INV (iii)
PB	time *t* (a)	Train	0.463	0.458	0.461
		Test	0.453	0.428	0.454
RF	time *t* (a)	Train	0.919	0.411	0.961
		Test	0.677	0.342	**0.699**
PB	dynamic (b)	Train	0.815	0.790	0.799
		Test	0.741	0.735	**0.745**
HRF	dynamic (b)	Train	0.915	0.748	0.901
		Test	**0.736**	0.552	0.733

to produce a fair comparison. Both PB and RF/HRF have been compared using PFI and SHAP values together with marginal effects. Exploring the change in probability associated with each predictor aimed at comprehending the impact on each class of the target variable. Simultaneously, more intricate relationships were evaluated using SHAP values. This process has facilitated the identification of the most significant features in terms of relative importance, which were then chosen for the implementation of the final rating model. The feature importance figures, referencing the top-performing statistical model (PB, dynamic caseversion) across the three variable sets, have been documented based on classification performance, as shown in Appendix 2.

Regarding PFI, the relative importance is calculated by taking the difference between the original and permuted F_1-score, followed by averaging and normalization over the sum of the absolute values of all obtained permutation metrics. This results in a scale ranging from 0% to 100%, with a negative score indicating that a random permutation of a feature's value leads to a better performance metric and a high importance score signifying that a feature is more sensitive to random shuffling, hence more "important" for prediction. In the selection of the most crucial predictors, features are evaluated individually based on their relative importance ranking. On an aggregated level, the total percentage of relative importance carried by the features in the top position is also considered. Related figures are presented on a macro-level (aggregated for all rating classes) and distinguished according to time dependence. The latter distinction has been carried out when both models (i.e., static and dynamic versions) report accuracy metrics higher than 50% on the test set. Otherwise, only one case has been analyzed.

PFI helps to easily make comparisons between features but it does not allow the proper assessment of the impact of features with medium permutation importance. Indeed, the Shapley explainer is crucial to correctly understand why a model predicts a given class for a given ID in a given time period (single row-prediction pair). SHAP goes through the input data, row-by-row and feature-by-feature, changing relative values to identify how much the base prediction differs, keeping fixed the rest for that row and, as a consequence, explaining how this prediction was reached. The Shapley value (*phi*) quantifies the contribution of each variable to the prediction of a single row, relative to the base prediction for the entire dataset. In a multiclass context, SHAP generates a distinct matrix for each class prediction for a given row, providing insights into

how each predictor influences the probability of belonging to that specific class, either increasing or decreasing it.

Figures 5, 6, and 7 report permutation importance metrics with reference to the PB model for the three sets of predictors. It can be stated that the autoregressive behavior seems to carry about 90% of relative importance on model prediction error. As a result, the other variables report negligible relative importance scores. Nonetheless, SHAP results allow to grasp individual contributions of variables on model predictions (Figs. 2, 3, and 4).

Starting from the FS set of variables, Fig. 2 highlights the high average contribution of Turnover, together with Current liabilities, Working capital, and EBIT. As expected, heterogeneous contribution is carried by the aforementioned features with respect to rating classes, with the highest impact on the highest ones (i.e., rating classes 4 and 5). EBIT and Working capital seem to have a significant effect also on the low-risk rating class 3.

Moreover, examining the marginal effects of the dynamic PB model for the set of variables based on financial statement (FS) allows for an analysis of how the probability changes when a predictor variable increases by one unit. According to Table 4, an increment in the key indicators reflecting the firm's financial solvency (i.e., Current liabilities, LT Debt, and Working capital) implies a positive impact on the probability of belonging to high-risk rating classes (classes 6 and 7). Increased long- and short-term financial obligations indicate greater debt and, consequently, higher risk. Similarly, working capital has similar effects on rating classes. As it is calculated as the difference between current assets and current liabilities (i.e., the sum of trade credit and payables), a positive sign may indicate a greater incidence of trade payables in the short term. In the case of small businesses, this can be particularly high and result in increased risk. Moreover, high working capital is identified as a potential indicator of liquidity tensions, suggesting that a company may not be efficiently reallocating capital for higher growth.

On the contrary, other variables, such as ROA, Liquid assets, Tangibles assets, Collections, and Turnover show negative marginal effects in correspondence with the riskiest rating classes. Given that high values for liquidity, profitability, and size measures indicate robust financial and operational performance, an increase in these metrics is correlated with a higher likelihood of belonging to low-risk rating classes. Specifically, high liquidity implies a better ability of the company to meet its short-term obligations on time, resulting in lower debt and, consequently lower risk. The annual sales volume is a signal of firm expansion and consolidated business model and is associated with a healthy corporate profile.

On the other side, SHAP results with reference to INV variables (Fig. 3) report a significant role of Outstanding on PB model predictions.

Delinquency and Outstanding represent metrics of economic exposure of the firms under investigation. The former reflects missed payments on invoices. According to our descriptive statistics, the average delinquency in our sample is 1.62%, with a standard deviation of around 10%. This appears in line with the statistics provided by the National Association of Factoring for the Italian Market (Assifact) which finds an average "past-due" of 1.18% over the period 2015–2022. The past due definition is aligned with our methodology that takes the exposure not paid within 90 days (Delinquency 90). Nevertheless, we do not apply Delinquency 90 because of the different rules on past due existing in Italy before 2017.[7] The latter represents the final debtor's economic exposure to invoice-related transactions. These metrics are directly linked to the level

[7] https://www.assifact.it/fact-news/il-mercato-del-factoring-parte-con-il-piede-giusto-nel-2023.

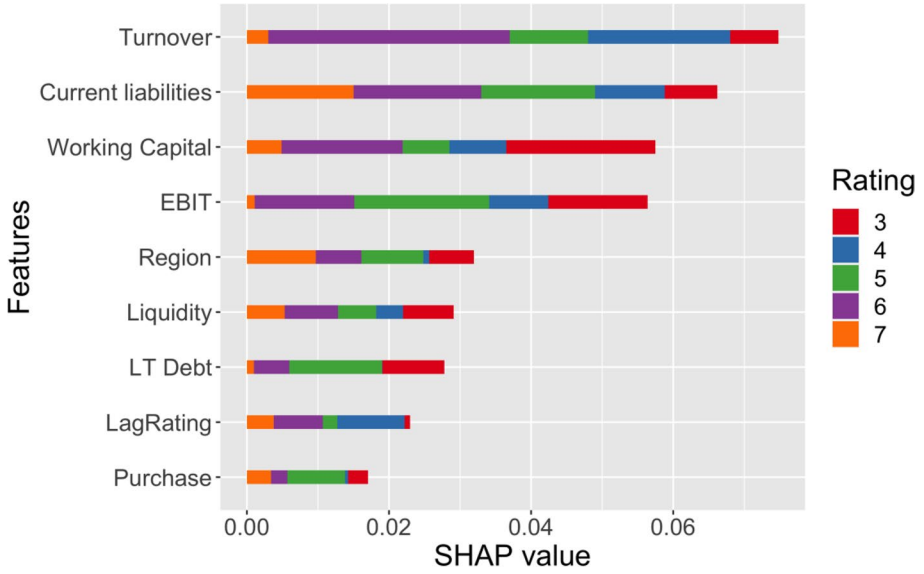

Fig. 2 SHAP value (average impact of predictors for each class) for the dynamic probit model with regards to FS set information pool (i), scenario (b)

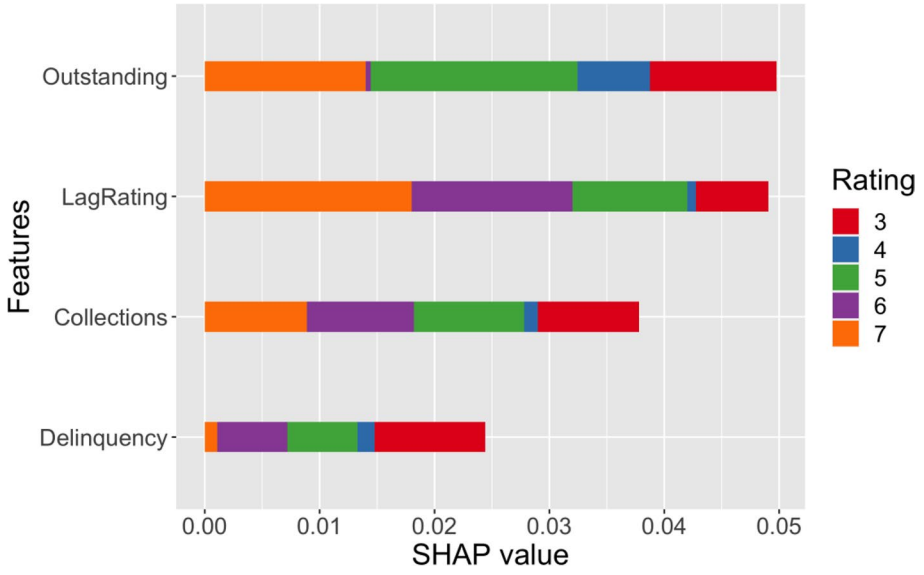

Fig. 3 SHAP value (average impact of predictors for each class) for the dynamic probit model with regards to INV set information pool (ii), scenario (b)

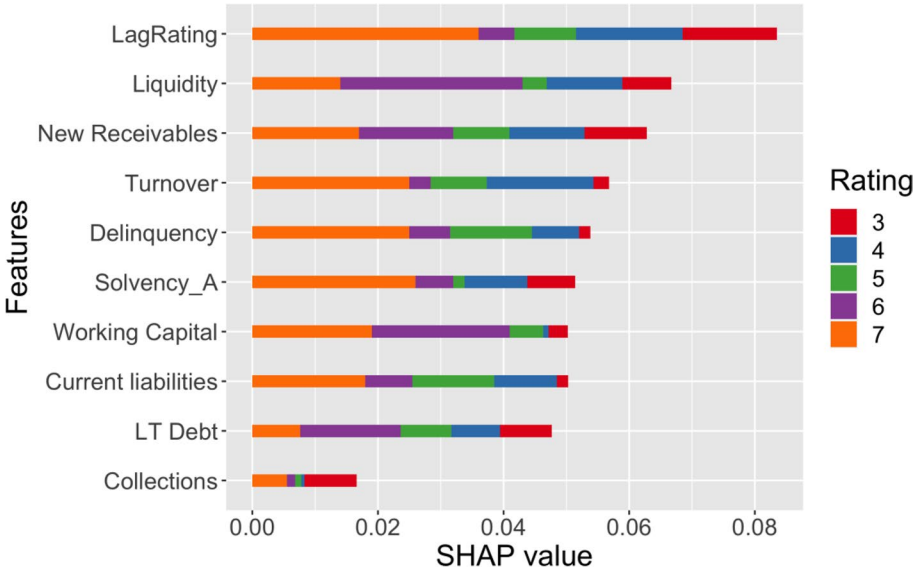

Fig. 4 SHAP value (average impact of predictors for each class) for the dynamic probit model with regards to FS + INV set information pool (iii), scenario (b)

of risk associated with each final debtor. The variable New Receivables instead measures the final debtor's trade credit position in terms of volume and lengthening of deadlines. A higher position of trade credit could reflect liquidity drainage, i.e., less financial resources readily available for investment (Filomeni et al. 2023b).

Tables 5 and 6 highlight that an increase in Delinquency results in a positive effect on the probability of belonging to the riskiest rating classes. The opposite behavior is shown by Outstanding and Collections.

Regarding the combined set of variables, Fig. 4 shows a high SHAP value for LagRating, immediately followed by Liquidity and New Receivables, showing the highest magnitude in terms of average impact by feature and class. It can be noticed that LagRating has the highest impact on the class with the highest risk (i.e., class 7), together with New Receivables, while Liquidity shows the highest average effect on rating class 6.

Specifically, the combination of LagRating, Turnover, Delinquency, and Solvency plays a significant role in the identification of the extreme rating class 7, bringing up or down the probability of belonging to that specific class.

4.2 Assessment of differences

The methodology outlined in Section 3.4 has been applied to conduct a statistical comparison of classifiers, aiming at evaluating significant differences in the results obtained in the preceding section. Initially, the macro-weighted balanced accuracy achieved by the aforementioned algorithms in the two distinct datasets (i.e., FS and INV) was imported and subsequent differences were examined on both the algorithm and the dataset levels. Since the results obtained from the Friedman test described in Section 3.4 show globally significant differences at the algorithm level, the next step involves

Table 4 Table of PB marginal effects for FS variables

Model	Historical	Variables	Marginal effects				
			y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static (scenario a)	Current liabilities	-0.2871 (****)	-0.5251 (****)	-0.1829 (****)	0.7660 (****)	0.2292 (****)
		Liquidity	1.1073 (ns)	2.0251 (ns)	0.7056 (ns)	-2.9543 (ns)	-0.8837 (***)
		LT Debt	-0.3299 (****)	-0.6034 (****)	-0.2102 (****)	0.8802 (***)	0.2633 (***)
		ROA	0.3262 (****)	0.5966 (****)	0.2079 (****)	-0.8702 (****)	-0.2603 (****)
		Tangibles	0.0233 (ns)	0.0425 (ns)	0.0148 (ns)	-0.0620 (ns)	-0.018 (ns)
		Working Capital	-0.0569 (****)	-0.1042 (****)	-0.0363 (****)	0.1520 (****)	0.0455 (****)
		Purchase	0.0131 (***)	0.0240 (**)	0.0083 (**)	-0.0349 (**)	-0.010 (**)
		Turnover	0.0584 (****)	0.1068 (****)	0.0372 (****)	-0.1558 (****)	-0.0466 (****)
		R1	0.0168 (*)	0.0295 (*)	0.0091 (*)	-0.0431 (*)	-0.0123 (*)
		R2	0.0057 (ns)	0.0102 (ns)	0.0033 (ns)	-0.0149 (ns)	-0.0043 (ns)
		R3	-0.0233 (**)	-0.0464 (**)	-0.0210 (*)	0.0675 (**)	0.0232 (**)
		N1	-0.0279 (ns)	-0.0540 (ns)	-0.0227 (ns)	0.0785 (ns)	0.0260 (ns)
		N2	-0.0225 (ns)	-0.0405 (ns)	-0.0136 (ns)	0.0591 (ns)	0.0175 (ns)
		N3	-0.0042 (ns)	-0.0079 (ns)	-0.0029 (ns)	0.0116 (ns)	0.8485 (ns)
		N4	0.0116 (ns)	0.0199 (ns)	0.0055 (ns)	-0.0291 (ns)	0.6509 (ns)
	Dynamic (scenario b)	Current liabilities	-0.0509 (****)	-0.3914 (****)	-0.2904 (****)	0.7069 (****)	0.0259 (****)
		Liquidity	0.2106 (ns)	1.6164 (ns)	1.1993 (ns)	-2.9192 (ns)	-0.1071 (ns)
		LT Debt	-0.0582 (****)	-0.4471 (****)	-0.3317 (****)	0.8074 (****)	0.0296 (****)
		Working Capital	-0.0114 (***)	-0.0878 (**)	-0.0651 (**)	0.1586 (**)	0.0058 (**)
		Purchase	0.0018 (*)	0.0142 (*)	0.0105 (*)	-0.0257 (*)	-0.0009 (*)
		EBIT	0.0409 (**)	0.3143 (**)	0.2332 (**)	-0.5676 (**)	-0.0208 (**)
		Turnover	0.0099 (****)	0.0757 (****)	0.0562 (****)	-0.1368 (****)	-0.005 (****)
		R1	0.0020 (ns)	0.0151 (ns)	0.0107 (ns)	-0.0269 (ns)	-0.0009 (ns)
		R2	0.0023 (ns)	0.0169 (ns)	0.0118 (ns)	-0.0299 (ns)	-0.0011 (ns)
		R3	-0.0034 (*)	-0.0274 (ns)	-0.0231 (ns)	0.0518 (ns)	0.0021 (ns)
		LagRating_4	-0.0157 (****)	-0.1522 (****)	-0.2306 (****)	0.3613 (****)	0.0371 (****)
		LagRating_5	-0.0314 (****)	-0.2405 (****)	-0.3465 (****)	0.5251 (****)	0.0933 (****)
		LagRating_6	-0.1026 (****)	-0.4008 (****)	-0.3555 (****)	0.6173 (****)	0.2417 (****)
		LagRating_7	-0.0264 (****)	-0.2293 (****)	-0.5007 (****)	-0.2166 (****)	0.9730 (****)

Table 5 Table of PB marginal effects for INV variables

Model	Version	Variables	Marginal effects				
			y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static (scenario a)	New Receivables	-0.052 (**)	-0.046 (**)	-0.0132 (**)	0.0727 (**)	0.0384 (**)
		Outstanding	0.0242 (****)	0.0217 (****)	0.0062 (****)	-0.0341 (****)	-0.0180 (****)
		Delinquency	-0.2319 (****)	-0.2080 (***)	-0.0593 (****)	0.3267 (****)	0.1725 (****)
	Dynamic (scenario b)	Collections	0.0042 (ns)	0.0164 (ns)	0.0084 (ns)	-0.0271 (ns)	-0.0021 (ns)
		Outstanding	0.0039 (****)	0.0158 (****)	0.0081 (****)	-0.0259 (****)	-0.0019 (****)
		Delinquency	-0.0542 (**)	-0.2143 (**)	-0.1103 (**)	0.3519 (**)	0.0269 (**)
		LagRating_4	-0.0342 (****)	-0.1718 (****)	-0.1851 (****)	0.3291 (****)	0.0621 (****)
		LagRating_5	-0.0656 (****)	-0.2709 (****)	-0.2998 (****)	0.4692 (****)	0.1670 (****)
		LagRating_6	-0.1835 (****)	-0.4003 (****)	-0.2866 (****)	0.4951 (****)	0.3755 (****)
		LagRating_7	-0.0541 (****)	-0.2563 (****)	-0.4291 (****)	-0.2386 (****)	0.9769 (****)

Table 6 Table of PB marginal effects for FBS + INV variables

Model	Historical	Variables	Marginal effects				
			y = 3	y = 4	y = 5	y = 6	y = 7
PB	Static (scenario a)	Delinquency	-0.1739 (****)	-0.2928 (****)	-0.0834 (****)	0.4206 (****)	0.1295 (****)
		Turnover	0.0659 (****)	0.1110 (****)	0.0316 (****)	-0.1594 (****)	-0.0491 (****)
		Working Capital	-0.0884 (****)	-0.1488 (****)	-0.0424 (****)	0.2138 (****)	0.0658 (****)
		LT Debt	-0.3978 (****)	-0.6697 (****)	-0.1908 (****)	0.9619 (****)	0.2963 (****)
		Current liabilities	-0.2948 (****)	-0.4963 (****)	-0.1414 (****)	0.7129 (****)	0.2196 (****)
		Liquidity	1.9809 (***)	3.3348 (***)	0.9503 (***)	-4.7904 (***)	-1.4757 (***)
	Dynamic (scenario b)	Delinquency	-0.0399 (**)	-0.2439 (**)	-0.1881 (**)	0.7069 (**)	0.4012 (**)
		Turnover	0.0141 (****)	0.0861 (****)	0.0482 (****)	-2.9192 (****)	-0.1416 (****)
		Working Capital	-0.0179 (****)	-0.1095 (****)	-0.0613 (****)	0.8074 (****)	0.1802 (****)
		LT Debt	-0.0850 (****)	-0.5190 (**)	-0.2905 (**)	0.1586 (**)	0.8537 (**)
		Current liabilities	-0.0639 (*)	-0.3902 (*)	-0.2184 (*)	0.0257 (*)	0.6418 (*)
		Liquidity	0.4997 (**)	3.0508 (**)	1.077 (**)	-0.5676 (**)	-5.0185 (**)
		LagRating_4	-0.0211 (****)	-0.1522 (****)	-0.1881 (****)	0.3613 (****)	0.0385 (****)
		LagRating_5	-0.0386 (****)	-0.2405 (****)	-0.2978 (****)	0.5251 (****)	0.0959 (****)
		LagRating_6	-0.1111 (****)	-0.4008 (****)	-0.3212 (****)	0.6173 (****)	0.2354 (****)
		LagRating_7	-0.0331 (****)	-0.2293 (****)	-0.4708 (****)	-0.2166 (****)	0.9537 (****)

Table 7 Corrected p-value matrix using Bergmann and Hommel's correction procedure generated when doing all the pairwise comparisons

		PB		HRF	
		time *t*	dynamic	time *t*	dynamic
PB	time *t*		**0.02**	0.52	0.21
	dynamic	**0.02**		0.12	0.52
HRF	time *t*	0.52	0.12		0.52
	dynamic	0.21	0.52	0.52	

analyzing those pairwise combinations that appear different. The related p-value matrix shows significant differences at 0.05 level for the PB model based on the different time dimension, thus highlighting the temporal component as a statistically significant discriminant between algorithms (Table 7).

5 Conclusions

Small businesses historically faced challenges in securing funding through traditional channels due to their limited size and information asymmetries. The emergence of Fintech has presented a solution to these obstacles, especially in the invoice lending environment. The rapid adoption of advanced techniques like machine learning (ML) by Fintech platforms raises transparency and reliability concerns. This paper fits in the literature on credit rating evaluation for SMEs and financial technologies applications to the environment of invoice lending by investigating the ability of ML techniques to correctly evaluate SMEs' creditworthiness. By exploiting a proprietary dataset of securitized invoices from Italian SMEs over the period 2015–2017, we evaluate the accuracy of ML algorithms compared with traditional probit models. Shapley values help interpret our ML's results. Our findings provide evidence of ML's efficacy, especially with limited information, thus advocating its inclusion in Fintech lending, benefitting both borrowers and lenders. This study therefore contributes to understand ML's role in digital financial markets, which, through the use of explainable AI tools, can improve transparency. Finally, the study compares ML with traditional models and shows its performance with respect to more consolidated approaches.

Our empirical results have important managerial and policy implications. Imprecise measurement of credit risk poses a potential threat to the stability of the financial sector, jeopardizing the crucial role that banks and lenders play in the economy. The use of fair, unbiased, correct, and explainable ML techniques can support borrowers to access funding, even when information is scant, without compromising the reliability related to the accuracy in corporate credit risk evaluation. This issue is particularly important in light of the small businesses' need for financing and the answer provided by the Fintech developments that bring new lenders and new forms of lending with timely and lean credit evaluation processes.

The study has some limitations that pave the way for further research on the topic. The period covered by our sample does not present specific macroeconomic shocks. Results could vary depending on the macroeconomic conditions, including financial crises, high inflation, and geopolitical risks. The evolving nature of ML methods can yield Fintech and traditional lenders to develop new solutions that can further improve the reliability and accuracy of algorithms, determining the superiority of ML methods over traditional ones (e.g., probit) even when complete information is available to the lender.

Appendix 1 Performance

Table 8 Model architecture for FS set of predictors information pool (i)

Model	Version	Scenario	Hyperparameters or Selected set of predictors
RF/HRF	Static	(a)	Mtry = 14; Ntrees = 500; Nodesize = 1
	Dynamic	(b)	Mtry = 6; Ntrees = 50; Nodesize = 3; Method = "meanw0"
PB	Static	(a)	Current liabilities + Liquidity ratio + LT Debt + ROA + Tangibles + Working Capital + Purchase + Turnover + Region + NACE
	Dynamic	(b)	Current liabilities + Liquidity + LT Debt + Working Capital + Purchase + EBIT + Turnover + Region + LagRating

Table 9 Model architecture for INV set of predictors information pool (ii)

Model	Version	Scenario	Hyperparameters or Selected set of predictors
RF/HRF	Static	(a)	Mtry = 5; Ntrees = 10; Nodesize = 100
	Dynamic	(b)	Mtry = 4; Ntrees = 141; Nodesize = 89; Method = "mean0"
PB	Static	(a)	New Receivables + Outstanding + Delinquency
	Dynamic	(b)	Collections + Outstanding + Delinquency + LagRating

Table 10 Model architecture for FS + INV set of predictors information pool (iii)

Model	Version	Scenario	Hyperparameters or Selected set of predictors
RF/HRF	Static	(a)	Mtry = 5; Ntrees = 500; Nodesize = 1
	Dynamic	(b)	Mtry = 5; Ntrees = 50; Nodesize = 3; Method = "freqw"
PB	Static	(a)	Collections + New Receivables + Delinquency + Turnover + Solvency + Working Capital + LT Debt + Current liabilities + Liquidity
	Dynamic	(b)	Collections + New Receivables + Delinquency + Turnover + Solvency + Working Capital + LT Debt + Current liabilities + Liquidity + LagRating

Appendix 2 Feature importance

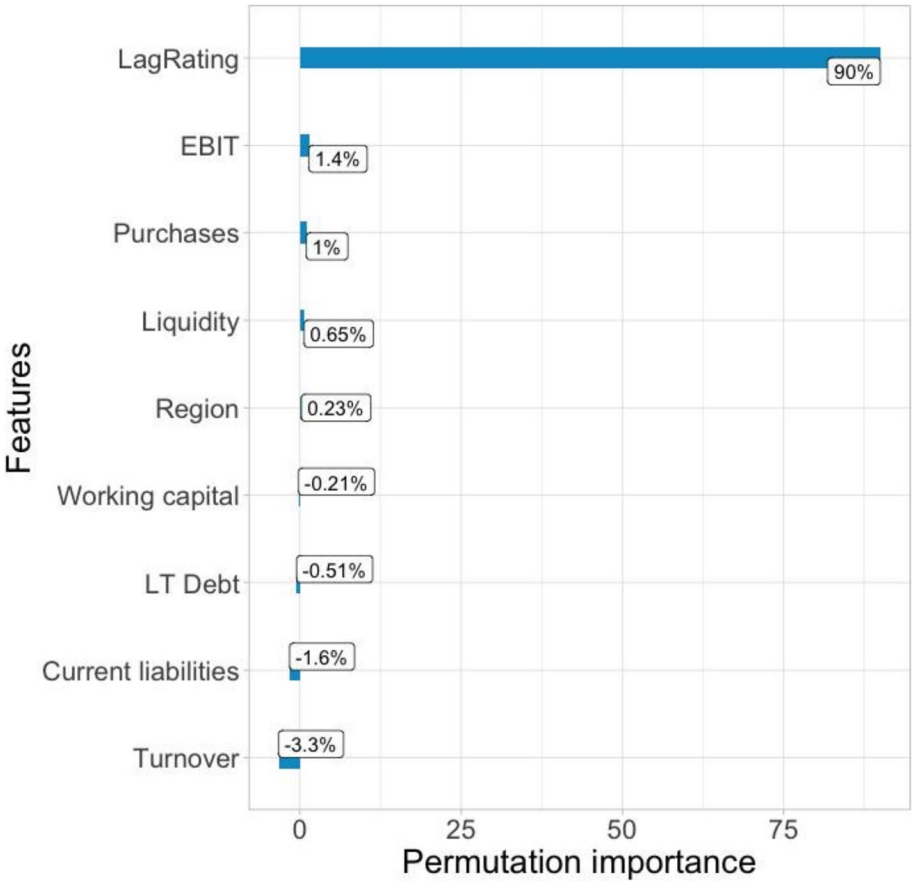

Fig. 5 Macro-averaged relative permutation importance for PB model for FS set (dynamic version) information pool (i), scenario (b)

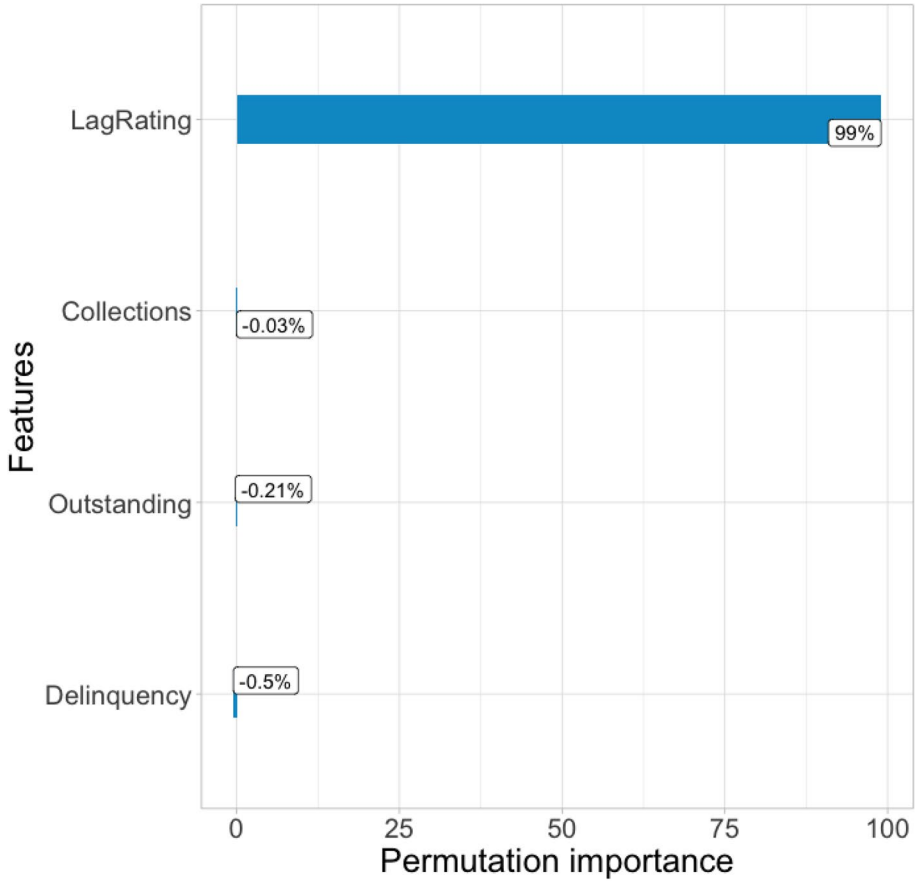

Fig. 6 Macro-averaged relative permutation importance for PB model for INV set (dynamic version) information pool (ii), scenario (b)

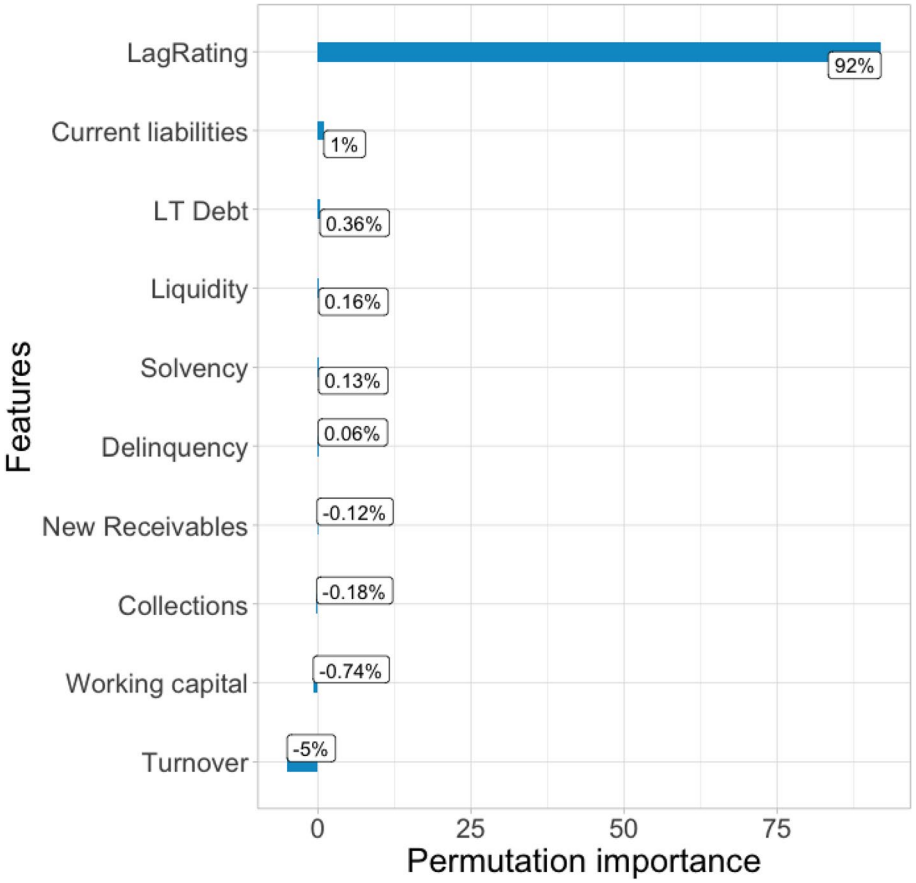

Fig. 7 Macro-averaged relative permutation importance for PB model for INV + FS set (dynamic version) information pool (iii), scenario (b)

Acknowledgements We are grateful to the Editor Cheng-Few Lee and to two anonymous referees for their valuable insights and suggestions.

Declarations

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abbasi K, Alam A, Brohi NA, Brohi IA, Nasim S (2021) P2p lending fintechs and SMEs' access to finance. Econ Lett 204:109890

Agostino M, Gagliardi F, Trivieri F (2012) Bank competition, lending relationships and firm default risk: An investigation of italian SMEs. Int Small Bus J 30(8):907–943

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Financ 23(4):589–609

Altman EI, Iwanicz-Drozdowska M, Laitinen EK, Suvas A (2017) Financial distress prediction in an international context: A review and empirical analysis of altman's z-score model. J Int Financ Manag Acc 28(2):131–171

Altman EI, Balzano M, Giannozzi A, Srhoj S (2023) The omega score: an improved tool for SME default predictions. J Int Council Small Bus 4(4):362–373. https://doi.org/10.1080/26437015.2023.2186284

Beaumont P, Tang H, Vansteenberghe E (2022) Collateral effects: the role of FinTech in small business lending. In: proceedings of the EUROFIDAI-ESSEC Paris December Finance Meeting

Beck T (2013) Bank financing for SMEs–lessons from the literature. Natl Inst Econ Rev 225(1):R23–R38

Berger AN (2006) Potential competitive effects of Basel II on banks in SME credit markets in the United States. J Financ Serv Res 29(1):5–36

Berger AN, Udell GF (1995) Relationship lending and lines of credit in small firm finance. J Bus 68(3):351–381

Berger AN, Udell GF (2006) A more complete conceptual framework for SME finance. J Bank Financ 30(11):2945–2966

Biau G, Scornet E (2016) A random forest guided tour. TEST 25:197–227

Bitetto A, Cerchiello P (2023) Initial coin offerings and ESG: allies or enemies? Fin Res Lett 57. https://doi.org/10.1016/j.frl.2023.104227

Bitetto A, Cerchiello P, Mertzanis C (2023) On the efficient synthesis of short financial time series: a dynamic factor model approach. Fin Res Lett 53. https://doi.org/10.1016/j.frl.2023.103678

Breeden J (2021) A survey of machine learning in credit risk. J Credit Risk 17(3):1–62

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

Bussmann N, Giudici P, Marinelli D, Papenbrock J (2021) Explainable machine learning in credit risk management. Comput Econ 57:203–216

Byanjankar A, Heikkilä M, Mezei J (2015) Predicting credit risk in peer-to-peer lending: a neural network approach. In 2015 IEEE symposium series on computational intelligence. IEEE, pp 719–725

Canales R, Nanda R (2012) A darker side to decentralized banks: market power and credit rationing in SME lending. J Financ Econ 105(2):353–366

Carroll N (2018) Estimation of ordered generalized linear models. https://CRAN.R-project.org/package=oglmx

Ciampi F, Giannozzi A, Marzi G, Altman EI (2021) Rethinking SME default prediction: a systematic literature review and future perspectives. Scientometrics 126(3):2141–2188

Claessens S, Krahnen J, Lang WW (2005) The Basel II reform and retail credit markets. J Financ Serv Res 28(1–3):5–13

Contoyannis P, Jones A, Rice N (2004) The dynamics of health in the British household panel survey. J Appl Economet 19:473–503

Dell'Ariccia G (2001) Asymmetric information and the structure of the banking industry. Eur Econ Rev 45(10):1957–1980

Dorfleitner G, Rad J, Weber M (2017) Pricing in the online invoice trading market: first empirical evidence. Econ Lett 161:56–61

Duarte FD, Gama APM, Gulamhussen MA (2018) Defaults in bank loans to SMEs during the financial crisis. Small Bus Econ 51(3):591–608

Filomeni S, Udell GF, Zazzaro A (2020) Communication frictions in banking organizations: evidence from credit score lending. Econ Lett 195C(109412). https://doi.org/10.1016/j.econlet.2020.109412

Filomeni S, Udell GF, Zazzaro A (2021) Hardening soft information: does organizational distance matter? Eur J Finance 27(9):897–927. https://doi.org/10.1080/1351847X.2020.1857812

Filomeni S, Bose U, Megaritis A, Triantafyllou A (2023a) Can market information outperform hard and soft information in predicting corporate defaults? Int J Financ Econ 1–26. https://doi.org/10.1002/ijfe.2840

Filomeni S, Modina M, Tabacco E (2023b) Trade credit and firm investments: empirical evidence from Italian cooperative banks. Rev Quant Financ Acc 60:1099–1141. https://doi.org/10.1007/s11156-022-01122-3

Filomeni S (2024) Securitization and risk appetite: empirical evidence from US banks. Rev Quant Fin Account Online First. https://doi.org/10.1007/s11156-024-01261-9

Financial Stability Board (2017) FinTechcredit: market structure, business models and financial stability implications. Comm Glob Financ Syst

Fisher A, Rudin C, Dominici F (2018) All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701. https://doi.org/10.1080/01621459.1937.10503522
Gomber P, Kauffman RJ, Parker C, Weber BW (2018) On the fintech revolution: Interpreting the forces of innovation, disruption, and transformation in financial services. J Manag Inf Syst 35(1):220–265
Gong C, Ribiere V (2021) Developing a unified definition of digital transformation. Technovation 102:102217
Gonzalez F, Haas F, Johannes R, Persson M, Toledo L, Violi R, Wieland M, Zins C (2004) Market dynamics associated with credit ratings. A literature review. Eur Central Bank Occas Paper 16:4–38
Gopal M, Schnabl P (2022) The rise of finance companies and fintech lenders in small business lending. Rev Financ Studies 35(11):4859–4901
Greene W, Hemsher D (2008) Modeling ordered choices: a primer and recent developments. Working Paper 26:1–181. New York University, Leonard N. Stern School of Business, Department of Economics
Grunert J, Norden L (2012) Bargaining power and information in SME lending. Small Bus Econ 39:401–417
Hadji-Misheva B, Osterrieder J (2023) A hypothesis on good practices for ai-based systems for financial time series forecasting: towards domain-driven xai methods. arXiv preprint arXiv:2311.07513
Hajjem A, Bellavance F, Larocque D (2014) Mixed-effects random forest for clustered data. J Stat Comput Simul 84:1313–1328
International Monetary Fund (2017) Fintech and financial services: initial considerations. IMF Staff Discussion Note 005:1–49
Ivashina V (2009) Asymmetric information effects on loan spreads. J Financ Econ 92(2):300–319
Kowalewski O, Pisany P (2022) The rise of fintech: a cross-country perspective. Technovation 122:102642
Liberti JM, Mian AR (2009) Estimating the effect of hierarchies on information use. Rev Financ Studies 22(10):4057–4090
Liberti JM, Petersen MA (2018) Information: hard and soft. Rev Corp Finance Studies 8(1):1–41
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:1–10
Medianovskyi K, Malakauskas A, Lakstutiene A, Yahia SB (2022) Interpretable machine learning for SME financial distress prediction. In international conference on computing and information technology. Springer, pp 454–464
Nguyen HH, Viviani J-L, Jabeur SB (2023) Bankruptcy prediction using machine learning and Shapley additive explanations. Rev Quant Fin Acc 1–42.https://doi.org/10.1007/s11156-023-01192-x
Odders-White E, Ready M (2006) Credit ratings and stock liquidity. Rev Financ Studies 19:119–157
OECD (2020) Financing SMEs and entrepreneurs: an OECD Scoreboard. Special edition: the impact of COVID-19. https://www.oecd.org/industry/smes/SMEs-Scoreboard-2020-Highlights-2020-FINAL.pdf
Ozili PK (2018) Impact of digital finance on financial inclusion and stability. Borsa Istanbul Rev 18(4):329–340
Schena C, Tanda A, Arlotta C, Potenza G (2018) The development of fintech. Consob FinTech Papers 1(March):15–122
Segal MR (1992) Tree-structured methods for longitudinal data. J Am Stat Assoc 87:407–418
Sela RJ, Simonoff JS (2012) RE-EM trees: a new data mining approach for longitudinal data. Mach Learn 86:169–207
Sexton J (2018) Historical tree ensembles for longitudinal data. https://CRAN.R-project.org/package=htree
Shapley LS (1953) A value for n-person games. Contrib Theory Games 2(28):307–317
Sharpe SA (1990) Asymmetric information, bank lending, and implicit contracts: a stylized model of customer relationships. J Financ 45(4):1069–1087
Soufani K (2002) On the determinants of factoring as a financing choice: evidence from the UK. J Econ Bus 54(2):239–252
Stein JC (2002) Information production and capital allocation: decentralized versus hierarchical firms. J Finance LVII(5):1891–1921
Strumbelj E, Kononenko I (2014) Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41(3):647–665
Tanda A, Schena C-M (2019) FinTech, BigTech and banks: digitalisation and its impact on banking business models. Springer
Thakor AV (2020) Fintech and banking: What do we know? J Financ Intermed 41:100833
The Royal Society (2019) Explainable AI: the basics. Available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf
Wooldridge J (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. J Appl Economet 20:39–54

Zhang BZ, Baeck P, Ziegler T, Bone J, Garvey K (2016) Pushing boundaries: The 2015 UK alternative finance industry report. https://ssrn.com/abstract=3621312