Annals of Operations Research
https://doi.org/10.1007/s10479-023-05306-5

S.I. : STATISTICAL METHODS AND DATA-DRIVEN FOR DECISION MAKING IN PUBLIC SECTOR

Websites' data: a new asset for enhancing credit risk modeling

Lisa Crosato[1] · Josep Domenech[2] · Caterina Liberati[3]

Accepted: 21 March 2023

Abstract
Recent literature shows an increasing interest in considering alternative sources of information for predicting Small and Medium Enterprises default. The usage of accounting indicators does not allow to completely overcome the information opacity that is one of the main barriers preventing these firms from accessing to credit. This complicates matters both for private lenders and for public institutions supporting policies. In this paper we propose corporate websites as an additional source of information, ready to be exploited in real-time. We also explore the joint use of online and offline data for enhancing correct prediction of default through a Kernel Discriminant Analysis, keeping the Logistic Regression and the Random Forests as benchmark. The obtained results shed light on the potentiality of these new data when accounting indicators lead to a wrong prediction.

Keywords Credit risk · SMEs · Web scraping · Corporate websites · Kernel discriminant analysis

1 Introduction

Small and Medium Enterprises (SMEs) characteristics, such as low value of assets to serve as collateral, a more rigid cost structure and limited cash reserves, make it harder for them

✉ Lisa Crosato
lisa.crosato@unive.it

Josep Domenech
jdomenech@upvnet.upv.es

Caterina Liberati
caterina.liberati@unimib.it

[1] Department of Economics, Ca' Foscari University of Venice and Bliss - Digital Impact Lab, Cannaregio 873, 30121 Venice, Italy

[2] Department of Economics and Social Sciences, Universitat Politècnica de València, Camí de Vera, s/n, 46022 Valencia, Spain

[3] Department of Economics, Management and Statistics and Center for European Studies (CefES), University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

Published online: 13 May 2023

to face abrupt crises (Belghitar et al., 2021) , either relying on their own forces or by accessing external financing, with respect to largest firms. Most of the times, information asymmetries play a key role in neglecting a credit line even to a potentially good firm since lenders cannot properly assess the worthiness of an SME, this being truer for the smallest, the most technologically advanced and the youngest companies (Martí and Quas, 2018) .

These difficulties should be circumvented since they can lead to barriers to growth, or, worst, to default (Cornille et al., 2019) . Indeed, preventing SMEs default, financing most promising firms and sustaining them in difficult times means protecting 99% of all enterprises in the EU, as well as the largest part of the European value added and jobs (56.4% and 66.6% respectively, European Commission (2019)). Large firms of course also rely on SMEs work and in some cases try to ease their way: for instance, established companies such as core retailers use their credit reputation to account for their small suppliers' credibility with banks (Zhu and Ou, 2021) .

Thus, it is not by chance that single governments and European Institutions promote SMEs-addressed support policies, such as the EU programme for the Competitiveness of Enterprises and Small and Medium-sized Enterprises (COSME), the Enterprise Europe Network (EEN), as well as non-specific support through the European Structural and Investment Funds (ESIF), to name a few (Padilla et al., 2018) .

According to empirical evidence (Cultrera, 2020; Dvouletỳ et al., 2021) , these funding programmes succeed by enhancing firm-survival, employment, fixed assets and sales, whereas the effects on labour productivity and total factor productivity are less clear-cut. All of these improvements are particularly strong for firms facing more severe financial constraints. An indirect effect of public funding is suggested by the certification hypothesis, implying that private investors may assess a firm quality by their succeeding in getting selective public support, since the latter already requires a thorough screening of firm characteristics (for a study on such effects in Spain, see Martí & Quas, 2018).

Information opacity, therefore, which marks especially the left side of the SMEs size distribution, is one of the main barriers preventing these firms from accessing to credit. This is one of the reasons why the literature is increasingly considering alternative or complementary sources of information with respect to accounting indicators, which are usually the basis for firms' screening. These sources comprehend legal judgments (Yin et al., 2020) , qualitative information on the management (Cornée, 2019) or on earnings management (Séverin and Veganzones, 2021) . The environmental context, like the institutional setup (Alexeev and Kim, 2012) or the local banking market (Arcuri and Levratto, 2020) was also taken into account. A new stream of the literature joins different kinds of data to collect all of the possible evidence for detecting firm default (Wang et al., 2022) .

A drawback of most of the above data is that they suffer of a large delay between their availability and their reference period. Furthermore, building up a prediction model accordingly requires getting access to them, which can be both costly and time consuming.

In this paper we propose corporate websites as an additional source of information for detection of SMEs default (Crosato et al., 2021) to be of help for avoiding both credit and public funding misallocation. On the one hand, web content data clearly require substantial efforts in data retrieval, selection, cleaning and ultimately analysis with respect to traditional sources of data. On the other hand, website information is free, assures the finest granularity, a large coverage of the firms' population and, most importantly, up-to-dateness. Previous works in the literature have already shown the usefulness of corporate websites to derive online proxies of firms' economic characteristics, such as corporate culture (Overbeeke and Snizek, 2005) , firm performance (Meroño-Cerdan and Soto-Acosta, 2007) , firm strategies (Llopis et al., 2010) or innovation (Axenbeck and Breithaupt, 2021) .

The online indicators can be obtained with an a priori classification of the web site contents or via web scraping of the corporate websites, as in Blazquez and Domenech (2018). The latter technique allows for an approach to monitoring based on the continuous observation of company behaviour, since changes or updates in firm websites might reveal credit-risk related variations in companies (Blazquez et al., 2018) . In this regard, the Wayback Machine of the Internet Archive is a valid tool to monitor websites as it offers a large repository with more than 26 years of web history.

Using about 900 Spanish SMEs sampled from the SABI[1] database, we build up a unique dataset combining the accounting (offline) indicators with our new online indicators. The joint use of online and offline information for enhancing correct prediction of default is explored through Kernel Discriminant Analysis (KDA, Baudat & Anouar, 2000), along with different mappings, keeping Logistic Regression (LR) and Random Forests (RF, Breiman, 2001) as benchmark. Separate predictions with either type of variable is also provided to assess whether and how the contribution of the online variables can be of help.

Results show that, albeit having a lower classification power, online indicators succeed in correctly identifying the future status of some firms whose accounting indicators would have led to the wrong prediction.

The paper is structured as follows: the next section sets out the methodology we have applied to transform websites into data and the websites characteristics that can be used to discriminate between surviving and defaulted firms, Sect. 3 briefly describes the dataset made by offline and online variables, Sect. 4 illustrates why we resort to non-linear discriminant analysis and the fundamentals of the method, Sect. 5 focuses on the results while Sect. 6 concludes.

2 Processing websites indicators

A website is a collection of linked documents stored in a web server. Business websites can be analyzed (or mined) from two different perspectives[2] (Liu and Chang, 2004) : web structure mining and web content mining. While the former focuses on how the different documents are linked, the latter concentrates on understanding the semantics and the meaning of the contents.

This section focuses only on content mining, since it is the closest approach to the business activity. Web content includes both the text and visuals that web browsers render, and also the code (generally in HTML and JavaScript) describing the layout, the organization and the interactivity with the user. Content mining is, thus, approached considering these two parts of the content: i) the mining of the text composing the business websites, and ii) the mining of the code describing the page. Code and textual content have been parsed to transform it into features (i.e., variables), which are are summarized in Table 1. These features will be used as input of the learning models. Mining images and videos is substantially more challenging, and has not been considered in this paper.

[1] SABI stands for Sistema de Análisis de Balances Ibéricos (Iberian Balance Sheets Analysis System). It is a database published by Bureau van Dijk Electronic Publishing (BvDP), a Moody's Analytics company.

[2] There is a third perspective (web usage mining) that can only be used by the owners of the website, which is not the case for the research presented in this paper.

Table 1 Feature types extracted from company websites

Feature type	Source	Description
Words	Text	Presence of a given word in the text
Stems	Text	Presence of a given stem in the text
Htmltags	Code	Presence of a given HTML tags in the code
Metaname	Code	Presence of a given metadata tag in the code
Linkhref_ext	Code	Presence of a given file extension in the Link tag
Hrefwords	Code	Presence of a given word in the URL of a Linked page
Href_ext	Code	Presence of a given file extension in the URL of a linked page

2.1 Textual content

Business websites are a reflection of the activity of the company (Blazquez and Domenech, 2014) and include a significant part of text. Therefore, it is expected that information such as the sector in which it operates and its market orientation (e.g., national or international, final consumer or other businesses) emerges from the analysis of this text. Changes in this text also mean that the company is changing its behaviour to some extent, and thus, it is still alive and investing in the website (Blazquez et al., 2018) .

The feature extraction from the website text can be done by means of general text mining and Natural Language Processing (NLP) techniques. They encompass a wide variety of processes for discovering information in textual data (Feldman and Sanger, 2007) .

The first step in NLP is tokenization, that is, dividing the text (represented as a sequence of characters) into basic units of content. The most simple approach for tokenization is considering that such units are the words composing the text. Another approach is stemming, which consists in transforming every word found to its stem (e.g., *industry* and *industrial* are reduced to *industri*). Since different words with the same stem are represented identically, stemming reduces the number of variables and the complexity of the data set.

Once tokens are defined, a simple but effective method to map a document into a fixed-length vector, the Bag-of-Words (BoW) model, is used. A vector $\mathbf{f} = (f_1, f_2, ..f_i, ..f_l)$ is assigned to each website text, where f_i denotes the frequency of the *i-th* token and l is the size of the collection of the tokens (Jones, 1972; Lan et al., 2008) . As we are interested only in the presence or absence of words or stems in a website, we further transform $\mathbf{f}$ into a vector of dummy entries, representing the presence of a token.

2.2 HTML code

The HTML language is a standard defined by the World Wide Web Consortium that describes the elements of a web page by using tags and their attributes. These tags are useful to describe the interaction (e.g., defining hyperlinks or forms), appearance (e.g., bold or italics), and structure (e.g., defining lists or different blocks) of a web page. Since it has evolved through the years, the tags used in a website may be related to how old its design is.

The way in which companies use HTML tags provides relevant information to capture the underlying behavior of companies. For instance, a FORM tag is usually employed to interact with the company/site. EMBED is generally employed to include Flash technology, which is currently being abandoned. LAYER is another legacy tag, used decades ago to design the web page (browsers support it for compatibility reasons). Similar to what has been done with the text content, HTML code has been tokenized by defining a token for each HTML tag appearing in the code.

Special attention must be put on the hyperlinks, which are defined in HTML with an A tag. Analyzing A tag attributes (particularly, href), allows us to detect connections with external agents and some website structure. This can be done by checking which text is included in the hyperlink reference, or href (e.g., twitter, government, associations), or which file extensions are used, as they are related to the underlying technology (e.g., php, asp, htm...) or to which information is offered (e.g., pdf, xls...). To account for this, both the words and the file extension in the "href" attribute are also extracted and defined as tokens.

Other tags that can provide additional information on the technologies used are META and LINK. The former informs about the page metadata, which may report how the website was generated, among other characteristics. The latter describes a connection to an auxiliary file, generally with additional code or style definitions. To transform these tags into features, the "name" attribute of META tags and the file extension in the "href" attribute of LINK tags are tokenized.

Afterward, all the tokens created from the HTML code are converted into dummies passing through the Bag-of-Words model.

3 Data description

In this work we merge offline and online data, coming from separate sources, referring to the period 2013-2015.

Offline data are retrieved from SABI, a large database including balance sheet data from 1.8 million companies in Spain and employment data of 1.34 million companies, 99.6% of which were SMEs. As is well known, defaulted firms generally represent the strict minority: referring to the time period covered by our analysis, only 2.1% of the SMEs were not active in 2015. The default rate remained low also within sectors, ranging from 1.1% in agriculture, forestry and fishing (section A in NACE rev.2) to 2.8% in food and accommodation service activities (section I in NACE rev.2).

In order to guarantee adequate variability within defaulted firms, we have drawn from SABI a convenience sample of 926 Spanish SMEs equally distributed between survived and defaulted companies.

Online data are web indicators automatically computed after scraping company websites, as discussed in detail in the previous section. They consist of 50 dummies selected out of more than ten thousand and represent the most discriminating features between failed and survived firms, as in Crosato et al. (2021). Afterwards, they are reduced to quantitative factors by means of a Multiple Correspondence Analysis (Benzécri, 1977; Greenacre, 1984) .

Our target variable indicates if the companies defaulted in 2015 ($y = 1$) or not ($y = 0$). As for explicative variables, the offline indicators (Number of Employees, Debt amount, Economic Profit and Productivity) refer to 2013, while websites' indicators refer to 2014. We consider data availability as if we were a public institution asking directly the firm for its last balance sheet. Hence, we are giving the offline variables a little advantage since in 2014 one could actually collect, from SABI, accounting indicators dated back to 2012.

The online data for 2014 were retrieved by accessing to the Wayback Machine of the Internet Archive, a well-known repository which tracks the WWW evolution over time.

4 Methods

Machine Learning literature is full of successful applications in the Credit Scoring domain. In such regard, the work of Baesens et al. (2003) compares the performance of various state-of-the-art classification algorithms (e.g. Logistic Regression, Discriminant Analysis, k-Nearest Neighbors) with Least-Square Support Vector Machines (LS-SVM, Suykens & Vandewalle, 1999), highlighting optimal performance of the latter in terms of overall prediction.

Accordingly, in this work we compare the performance of the Kernel Discriminant Analysis, which can be reformulated into LS-SVM (Liberati et al., 2017) , with two benchmark algorithms. The first one is Logistic Regression which is still accounted as the industry standard, the second are Random Forests, considered in the recent literature the reference model for classification algorithms (Lessmann et al., 2015) .

Kernel-based methods are classification techniques based on statistical learning theory (Vapnik, 1995, 1998) . They convert a non linear problem into a linear one by projecting the data onto a high dimensional Feature Space $\mathcal{F}$. Clearly, a high or infinite dimension of $\mathcal{F}$ will make it impossible to directly solve the map. In these cases it is more efficient to seek a formulation of the algorithm which uses only dot-products of the training patterns. Kernels are functions $k : \mathcal{X}^2 \rightarrow \mathbb{R}$ which for all pattern sets $\{x_1, x_2..x_n\} \subset \mathcal{X}$ and with $\mathcal{X} \subset \mathbb{R}^p$, lead to positive matrices $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ (Schölkopf et al., 1999) . If the Mercer's theorem holds (Mercer, 1909) , then the kernel k gives rise to a (usually non linear) map $\phi : \mathbb{R}^p \rightarrow \mathcal{F}$ based on dot product of the pattern sets (Vapnik, 1995) , i.e.

$$k(\mathbf{x}, \mathbf{z}) = (\phi(\mathbf{x}) \cdot \phi(\mathbf{z})) \tag{1}$$

A Fisher Discriminant Analysis can then be carried out on the mapped data.

Putting it in a general prediction framework, let us consider the input dataset $\mathcal{I}_{XY} = \{(x_1, y_1), ..., (x_n, y_n)\}$ composed by the training vectors $x_i \in \mathcal{X}$ and the corresponding label values $y_i \in \mathcal{Y} = \{1, 0\}$. The class separability in a direction of the weights $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_n]'$ is obtained by maximizing the Rayleigh coefficient:

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}' \mathbf{S}_B^{\Phi} \boldsymbol{\alpha}}{\boldsymbol{\alpha}' \mathbf{S}_W^{\Phi} \boldsymbol{\alpha}} \tag{2}$$

where $\mathbf{S}_B^{\Phi}, \mathbf{S}_W^{\Phi}$ are the Between and Within covariance matrices in the Feature Space (Mika et al., 1999; Schölkopf et al., 1999) , respectively. This problem can be solved by finding the leading eigenvectors of $(\mathbf{S}_W^{\Phi})^{-1}\mathbf{S}_B^{\Phi}$. Since the proposed setting is ill-posed because the Within covariance matrix is at most of rank $n - 1$, we stabilized the matrix according to Thomaz et al. (2004) and we then employed the convex sum covariance estimator introduced by Chen (1976). For more details about Covariance Estimators see Pamukçu et al. (2015).

As a consequence, the kernel discriminant function $f(\mathbf{x})$ of the binary classifier can be written as

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(x_i, \mathbf{x}) \tag{3}$$

Finally, the group membership (g) of a new instance $\mathbf{x}_0$ is obtained in two steps: first computing the projection of the instance into the kernel discriminant space:

$$\widehat{y}_0 = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}_0) \tag{4}$$

Table 2 Kernel functions

Kernel mapping	k(**x**,**z**)
Cauchy (CAU)	$\frac{1}{1+\frac{\mathbf{x}-\mathbf{z}^2}{c}}$
Laplace (LAP)	$\exp(-\sqrt{\frac{\|\mathbf{x}-\mathbf{z}\|^2}{c^2}})$
Multi-quadric (MULTIQ)	$\sqrt{\mathbf{x}-\mathbf{z}^2+c^2}$
Polynomial (POLY)	$(\mathbf{x}\cdot\mathbf{z})^d$
Gaussian (RBF)	$\exp(\frac{-\|\mathbf{x}-\mathbf{z}\|^2}{2c^2})$

and, second, allocating the observation to the group whose centroid ($\overline{y}_g$) is the closest to its projection:

$$\text{cutoff} = \arg\min_g \|\widehat{y}_0 - \overline{y}_g\| \quad g = 1, 2 \tag{5}$$

5 Results

Our empirical analysis proposes various KDA solutions to be compared against the two bechmark models; we thus employ 5 different kernels among the most commonly used (Table 2). A grid search is used to set the width value, c, whereas the degree of the Polynomial kernel is fixed to $d = 2, 3$.

The classification protocol randomly splits the data into training (70%) and test (30%) set. To get robust classification rates, we generate 100 random samples and assess the performances of the competing classifiers by comparing the average values of the AUC (Area Under the receiver operating characteristic Curve), generally used as a reference metric in the literature. Since we are interested in the effectiveness of our solutions in both groups, we also monitor average values of sensitivity (correct default prediction), specificity (correct survived prediction) and total error (percentage of total misclassified instances).

We articulate the analysis in three settings, based on online only, offline only and both offline and online indicators. Performance metrics on the test set are collected in Tables 3 - 4 - 5 respectively.

A general inspection of the results reveals that online information has good predicting power in recognizing companies' bankruptcy, although the classification rates are not optimal. Most of the times, the usage of the KDA increases the AUC values with respect to the LR and the RF. In particular, using the online variables (Table 3), the best solution is provided by the Multi-quadric kernel ($c = 2.000$) with $AUC = 0.710$.[3]

Considering the offline variables, as in the former case, the non linear discriminants improve the classification performance: the RF and almost all the kernel-based classifiers provide AUCs higher than the 0.814 scored by the LR model (Table 4). Specifically, both the Laplace kernel ($c = 2.667$) and the RF reach the highest value of AUC (0.847) and the lowest error rate (0.228). Although the two classifications are equivalent according to our global performance metrics, the Laplace-KDA provides an optimal combination of sensitivity (0.772) and specificity (0.772) rates with respect to the RF which shows just a fair detection level of defaulted companies (sensitivity 0.754).

[3] In case of equal AUC, we select the model with the lowest Error rate.

Table 3 Test set metrics using online data: average values on 100 runs

c values	0.050	0.334	0.667	1.001	1.334	1.667	2.000	2.334	2.667	3.000
RBF										
Total error	0.499	0.436	0.376	0.356	0.354	0.352	0.353	0.354	0.355	0.355
Specificity	0.108	0.794	0.720	0.687	0.674	0.666	0.657	0.650	0.644	0.643
Sensitivity	0.894	0.335	0.529	0.601	0.619	0.630	0.637	0.642	0.647	0.647
AUC	0.518	0.597	0.668	0.699	0.706	0.709	0.709	0.709	0.710	0.710
LAP										
Total error	0.486	0.430	0.377	0.361	0.356	0.354	0.354	0.353	0.353	0.353
Specificity	0.603	0.714	0.707	0.705	0.702	0.699	0.696	0.695	0.694	0.693
Sensitivity	0.425	0.425	0.539	0.574	0.587	0.594	0.596	0.599	0.601	0.602
AUC	0.561	0.614	0.661	0.682	0.691	0.696	0.698	0.700	0.701	0.702
CAU										
Total error	0.436	0.381	0.365	0.360	0.357	0.357	0.356	0.356	0.355	0.355
Specificity	0.705	0.709	0.708	0.702	0.699	0.689	0.686	0.684	0.681	0.680
Sensitivity	0.423	0.529	0.561	0.577	0.586	0.597	0.602	0.605	0.609	0.610
AUC	0.600	0.656	0.678	0.688	0.693	0.697	0.699	0.701	0.702	0.704
MULTIQ										
Total error	0.350	0.353	0.353	0.352	0.354	0.353	0.354	0.355	0.355	0.355
Specificity	0.685	0.670	0.665	0.665	0.657	0.657	0.647	0.643	0.643	0.642
Sensitivity	0.614	0.624	0.628	0.630	0.635	0.637	0.645	0.648	0.648	0.649
AUC	0.705	0.707	0.708	0.709	0.709	0.709	0.710	0.709	0.710	0.710
POLY	degree 2	degree 3								
Total error	0.370	0.396								
Specificity	0.616	0.569								
Sensitivity	0.643	0.693								
AUC	0.692	0.695								
	LR	*RF*								
Total error	0.368	0.378								
Specificity	0.637	0.610								
Sensitivity	0.626	0.634								
AUC	0.696	0.657								

Combining online with offline variables (see Table 5) implies a moderate increase in terms of global classification (AUC), for a maximum of 0.9% (Laplace kernel with $c = 2.334$) with respect to the best score reached with the same classifier but using only the offline dataset. On the same time, in this setting we gain more than 5.4% in specificity although facing a little drop in sensitivity (less than 0.65%).

Concluding, offline variables display a higher classification power with respect to the online indicators. However, there is still room for correct classification of further firms, especially among the survived, by joining them together. We think it is worth to investigate if there are any peculiarities characterizing the firms classified according to the three variables settings of Tables 3 - 4 - 5, to understand more specifically the features workings.

Table 4 Test set metrics using offline data: average values on 100 runs

c values	0.050	0.334	0.667	1.001	1.334	1.667	2.000	2.334	2.667	3.000
RBF										
Total error	0.464	0.414	0.334	0.256	0.235	0.232	0.234	0.236	0.237	0.236
Specificity	0.683	0.344	0.601	0.744	0.764	0.767	0.757	0.746	0.741	0.739
Sensitivity	0.389	0.827	0.732	0.745	0.765	0.768	0.774	0.782	0.786	0.789
AUC	0.679	0.730	0.779	0.807	0.828	0.837	0.839	0.840	0.840	0.841
LAP										
Total error	0.446	0.354	0.270	0.240	0.234	0.230	0.229	0.228	0.228	0.228
Specificity	0.658	0.643	0.724	0.738	0.758	0.766	0.770	0.773	0.772	0.771
Sensitivity	0.450	0.649	0.736	0.782	0.773	0.773	0.773	0.771	0.772	0.772
AUC	0.714	0.767	0.801	0.832	0.839	0.844	0.846	0.846	0.847	0.847
CAU										
Total error	0.269	0.248	0.242	0.239	0.236	0.233	0.232	0.231	0.230	0.229
Specificity	0.685	0.702	0.723	0.738	0.747	0.755	0.760	0.764	0.767	0.769
Sensitivity	0.778	0.802	0.794	0.785	0.782	0.779	0.776	0.775	0.774	0.773
AUC	0.804	0.827	0.833	0.838	0.840	0.841	0.842	0.842	0.843	0.843
MULTIQ										
Total error	0.240	0.241	0.241	0.242	0.243	0.244	0.245	0.246	0.248	0.249
Specificity	0.727	0.726	0.724	0.721	0.717	0.715	0.712	0.707	0.702	0.696
Sensitivity	0.792	0.793	0.793	0.794	0.796	0.797	0.799	0.801	0.803	0.806
AUC	0.840	0.839	0.838	0.837	0.837	0.836	0.835	0.834	0.834	0.833
POLY	degree 2	degree 3								
Total error	0.377	0.481								
Specificity	0.552	0.917								
Sensitivity	0.694	0.121								
AUC	0.786	0.648								
	LR	*RF*								
Total error	0.258	0.228								
Specificity	0.729	0.790								
Sensitivity	0.755	0.754								
AUC	0.814	0.847								

5.1 Firm-by-firm analysis

In the following, we concentrate on the best performing model in each of the three settings[4] with a view to study the percentage of correct classification for each firm across simulations. Thus, we consider the companies who are correctly classified by:

I) The offline variables
II) The combined variables
III) The online variables but not by the offline variables

[4] Following results in Table 4, the best model for the offline variables setting is the Laplace-KDA.

Table 5 Test set metrics combining online and offline data: average values on 100 runs

c values	0.050	0.334	0.667	1.001	1.334	1.667	2.000	2.334	2.667	3.000
RBF										
Total error	0.465	0.334	0.241	0.218	0.215	0.217	0.221	0.227	0.231	0.233
Specificity	0.656	0.859	0.851	0.823	0.800	0.789	0.773	0.750	0.736	0.728
Sensitivity	0.413	0.472	0.668	0.741	0.769	0.777	0.786	0.796	0.802	0.805
AUC	0.709	0.787	0.822	0.839	0.847	0.850	0.850	0.847	0.846	0.844
LAP										
Total error	0.445	0.282	0.221	0.211	0.210	0.210	0.210	0.209	0.210	0.210
Specificity	0.706	0.806	0.804	0.814	0.816	0.816	0.814	0.814	0.813	0.810
Sensitivity	0.404	0.631	0.753	0.764	0.763	0.764	0.766	0.767	0.768	0.771
AUC	0.749	0.791	0.835	0.847	0.852	0.853	0.855	0.855	0.855	0.854
CAU										
Total error	0.240	0.213	0.210	0.210	0.210	0.210	0.210	0.210	0.210	0.210
Specificity	0.778	0.803	0.814	0.816	0.817	0.816	0.814	0.813	0.812	0.810
Sensitivity	0.742	0.770	0.766	0.765	0.764	0.765	0.766	0.767	0.769	0.770
AUC	0.819	0.84	0.845	0.849	0.851	0.852	0.852	0.852	0.853	0.853
MULTIQ										
Total error	0.228	0.231	0.233	0.235	0.238	0.240	0.243	0.244	0.246	0.249
Specificity	0.745	0.741	0.733	0.726	0.717	0.708	0.701	0.694	0.686	0.679
Sensitivity	0.798	0.798	0.800	0.804	0.807	0.811	0.814	0.818	0.821	0.824
AUC	0.840	0.839	0.839	0.839	0.838	0.838	0.837	0.836	0.835	0.835
POLY	degree 2	degree 3								
Total error	0.404	0.484								
Specificity	0.535	0.912								
Sensitivity	0.656	0.121								
AUC	0.705	0.638								
	LR	*RF*								
Total error	0.251	0.235								
Specificity	0.733	0.798								
Sensitivity	0.764	0.733								
AUC	0.819	0.843								

IV) The combined variables but not by the offline variables

in more than 50% of the samples in which they are included.

Table 6 reports the medians of the offline variables calculated separately for defaulted and survived firms as well as the p.value of the Wilcoxon rank test for the difference in medians, in cases (I) to (IV) and with respect to the observed dependent variable.

It comes at no surprise that the accounting indicators classification reflects what is commonly observed: failed firms have, on average, a smaller number of employees, higher debt, negative profits and lower productivity with respect to survived. But in this dataset, as well as in others and although less frequently, there are also firms who fail even if they are not the smallest or if part of the reported accounts are in a good state: that is where the online

Table 6 Median values for defaulted (D) and survived (S) firms as observed and as identified by different types of variables; pvalue of the Wilcoxon rank test

Variables		Observed data		Detected by offline		Combined		Online only		Combined only	
				(I)		(II)		(III)		(IV)	
		Med	pvalue	Med	pvalue	Med	pvalue	Med	pvalue	Med	pvalue
Employees	D	8.00	0.000	5.00	0.000	5.00	0.000	56.00	0.000	9.00	0.000
	S	35.00		42.00		41.00		10.00		24.50	
Debt	D	75.00	0.000	79.50	0.000	79.60	0.000	68.30	0.837	72.80	0.005
	S	55.10		54.90		54.10		68.20		36.00	
Economic profit	D	−0.50	0.000	−2.36	0.000	−2.30	0.000	0.90	0.287	0.03	0.054
	S	1.50		1.89		−2.00		0.85		4.80	
Productivity	D	29.60	0.000	27.00	0.000	26.70	0.000	40.40	0.237	40.10	0.355
	S	46.40		47.50		47.80		47.10		47.40	

Table 7 Classification metrics calculated as: number of firms correctly identified in more than 50% of the samples over the total number of observed instances in either group for defaulted (D) and survived (S) firms as identified by different types of variables

	Offline (I)	Combined (II)	Online only (III)	Combined only (IV)
D	76.6%	76.8%	8.5%	2.1%
S	78.8%	82.3%	10.6%	5.2%

indicators can supply additional information. We can see that online indicators, in fact, are able to spot these accounting-anomalous firms, which count on average more employees for failed with respect to survived together with similar levels of Debt and Economic Profit. Likewise, the difference in median Productivity between the two groups rightly identified by the online features is reduced -and not significant- as compared with the same difference between the observed groups (Table 6).

To measure the online features' contribution to the classification, we have recalculated the metrics in either group as the number of firms correctly identified in more than 50% of the samples over the total number of instances. As can be seen from Table 7, percentages of correct classification in both groups are very similar to those achieved in Tables 4 and 5. But, if we consider those firms correctly predicted by online indicators and mispredicted by offline variables, we can regain 8.5% of the defaults and 10.6% of the survivors.

Unfortunately, the additional information carried by the online indicators gets in part lost when combining them to the offline ones, so that the recovery is less pronounced (2.1% and 5.2% respectively). This loss is probably to be ascribed to the largest discriminating power between the two groups provided by the accounting indicators that, as a consequence, pull the classification on their side. This analysis of course calls for alternative approaches on the joint use of various types of variables, but on the same time stresses the important point that online and offline indicators can be complementary tools for firms' evaluation because they identify different firms.

As about the presence of particular features in the websites, the assessment of significant differences between surviving and defaulted firms is complicated by the small number of occurrences resulting for some features.

To account for this issue we calculate two alternative two-sided 10% confidence intervals, namely the Newcombe Score method (Newcombe, 1998) and the method proposed by Agresti and Caffo (2000), along with the classic test for difference in proportions (Table 8). The direction for the alternative hypothesis in the latter, whether an indicator is more present among the survived or the defaulted, is formulated according to what suggested by the estimated intervals.

Among the features more commonly discriminating towards surviving firms, some describe that the companies own resources, such as "words_propias" (Spanish for *own*) and "words_maquinaria" (Spanish for *equipment*), as well as offering virtual resources for downloading ("hrefwords_descargas" represents a link to download resources). It is also noticeable that offering recent news (the "stems_sal" feature is related to a press room section), a wide range of products (related to "stems_gam"), and working in the industrial sector (related to "words_industrial") is connected to higher survival rates.

Websites belonging to healthy companies generally display the street ("stem_carreter") where the firms are located, as well as the possibility to mirror the contents in different languages ("stems_castellan" is connected to the name of the Spanish language, while "stems_cat" is short for Catalan). In summary, it is more likely that surviving firms organize

Table 8 Tests for difference in the presence of online indicators between survived and defaulted firms when detected by online indicators only: *p*values of the Wald test, Agresti-Caffo and Newcombe Score 10% confidence intervals

Indicator	Agresti-Caffo	Newcombe score	*p*value
Highest presence in survived			
Stems_sal	0.032, 0.224	0.039, 0.232	0.015
Words_propias	0.015, 0.157	0.026, 0.170	0.019
Words_maquinaria	0.015, 0.157	0.026, 0.170	0.019
Stems_carreter	0.014, 0.173	0.023, 0.183	0.028
Stems_cat	0.001, 0.131	0.012, 0.144	0.051
Words_industrial	0.001, 0.131	0.012, 0.144	0.051
Hrefwords_descargas	0.001, 0.131	0.012, 0.144	0.051
Stems_castellan	0.001, 0.131	0.012, 0.144	0.051
Stems_gam	0.002, 0.190	0.008, 0.197	0.066
Highest presence in defaulted			
Stems_es	– 0.441, – 0.166	– 0.440, – 0.167	0.001
Metaname_generator	– 0.347, – 0.111	– 0.348, – 0.112	0.003
Linkhref_ext_php	– 0.199, – 0.039	– 0.208, – 0.044	0.019
Stems_con	– 0.308, – 0.031	– 0.307, – 0.031	0.035
Words_qué	– 0.141, – 0.015	– 0.154, – 0.022	0.042
Hrefwords_info	– 0.198, – 0.011	– 0.202, – 0.012	0.056
Words_nuestro	– 0.210, – 0.009	– 0.212, – 0.010	0.060

their sites in such a way to facilitate the web-surfing among the information: words and features help to better direct the access and research of the internauts.

On the contrary, the features that discriminate towards defaulted firms describe some limitations in their resources. For instance, “metaname_generator” point out the use of some Content Management System (CMS), such as Wordpress, which are inexpensive solutions for publishing a website. This can also be the reason behind “linkhref_ext_php” because CMSs tend to use PHP technology. Other features in this group are related to the way companies appeal to website visitors. For instance, “words_qué” is a feature present in the sentences *¿Qué hacemos?* (What do we do?) and *¿Por qué nosotros?* (Why us?), commonly used as section titles. This wording could be perceived as an informal or homespun communication style. The same informal style is behind “words_nuestro” (Spanish for *our*), or “hrefwords_info”, which points out a website section including generic information. Finally, the websites show references to the Spanish language (“stem_es”): since the analysed sample is composed by Spanish companies, we cannot consider such trait as an active behaviour of the firms to enlarge the pool of their potential customers. Therefore, websites of defaulted firms seem to be less external communication oriented, appearing sloppy and poorly structured.

6 Conclusion

Alternative sources of data have become increasingly important for the study of SMEs, for both financial intermediaries and public institutions.

Our paper contributes to this field by proposing the use of website data to enhance credit risk modelling in a balanced defaulted/surviving firm setting.

Results show that KDA, particularly with Laplace and Multiquadric kernels, delivers the best classification metrics, although Random Forests are a suitable alternative on traditional indicators, less on the unconventional data-source proposed in this paper. We also find that offline indicators got the better of online ones, leading to a smallest global classification error. However, online features do reveal a certain predictive power that, together with their free and real-time availability, represents a non-negligible asset for assessing a firm's creditworthiness, as shown by the increased AUC provided by the combination of the two types of data.

Another contribution of the paper is to concentrate on the pitfalls of the accounting indicators, an aspect which, to the best of our knowledge, was not investigated in the past literature. Considering a largest sample of defaulted firms allowed a more thorough examination of the cases where accounting indicators may lead to a wrong prediction and where, on the contrary, alternative data can make the difference. Our analysis confirms that models based on accounting measures do their job in correctly predicting default, as long as the firm accounting profile is in line with the expectations: good figures for good firms and bad figures for bad firms. But what about companies who are not in line with these expectations? They could be simply undergoing a serious downturn but be resilient enough to overcome it, or they could be "zombies": companies who have been in financial distress for several years, but are still in the business also due to external funding (Blažková & Dvouletỳ 2022). On the opposite, they could be cheating on their figures to obtain funds.

Our assessment at the firm-level attests that indicators scraped from firms' websites could help in predicting whether a firm will make it or not, and thus whether it can be a horse to bet on for public support, particularly when accounting figures would have led to misclassification.

Finally, we also have described both defaulted and surviving firms according to the features present in their websites, providing evidence that the care and time companies spend on them are a proxy of their future health.

This work does have some limitations that further research should attempt to overcome. Trying different modelling designs for combining the offline and online variables to exploit at a maximum the different informative content of both, as well as investigating more thoroughly the website features revealing the state of a firm are essential. Additionally, although accounting indicators are a pillar for assigning public funds, more research should focus on alternative sources of evaluation to characterize companies whose balance sheets can lead to misallocation.

Acknowledgements The authors thank two anonymous referees for their helpful comments.

Funding This work was supported by grant PID2019-107765RB-I00 (funded by MCIN/AEI/10.13039/501100011033). The Open Access publishing option was funded by the Ca' Foscari University of Venice, Italy.

Declarations

Conflict of interest The authors have no conflicts of interest to declare.

References

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician, 54*(4), 280–288.

Alexeev, M., & Kim, J. (2012). Bankruptcy and institutions. *Economics Letters, 117*(3), 676–678.

Arcuri, G., & Levratto, N. (2020). Early stage SME bankruptcy: does the local banking market matter? *Small Business Economics, 54*(2), 421–436.

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites-which website characteristics predict firm-level innovation activity? *PloS One, 16*(4), e0249583.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society., 54*, 627–635.

Baudat, G., & Anouar, F. (2000). Generalized discriminant analysis using a kernel approach. *Neural Computation, 12*, 2385–2404.

Belghitar, Y., Moro, A., & Radić, N. (2021). When the rainy day is the worst hurricane ever: the effects of governmental policies on SMEs during COVID-19.*Small Business Economics*1–19.

Benzécri, J. P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Les Cahiers de l'Analyse des Données, 2*, 55–71.

Blažková, I., & Dvouletỳ, O. (2022). Zombies: Who are they and how do firms become zombies? *Journal of Small Business Management, 60*(1), 119–145.

Blazquez, D., & Domenech, J. (2014). Inferring export orientation from corporate websites. *Applied Economics Letters, 21*(7), 509–512.

Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy, 24*(2), 406–428.

Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites changes reflect firms survival? *Online Information Review, 42*(6), 956–970.

Breiman, L. (2001). *Random forests. Machine learning, 4*(5), 15–32.

Chen, M.C.F.(1976). *Estimation of covariance matrices under a quadratic loss function*. Research Report S-46.Department of Mathematics SUNY at Albany, Albany, N.Y.

Cornée, S. (2019). The relevance of soft information for predicting small business credit default: Evidence from a social bank. *Journal of Small Business Management, 57*(3), 699–719.

Cornille, D., Rycx, F., & Tojerow, I. (2019). Heterogeneous effects of credit constraints on SMEs' employment: Evidence from the European sovereign debt crisis. *Journal of Financial Stability, 4*, 11–13.

Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting SME's default: Are their websites informative? *Economics Letters, 204*, 109888.

Cultrera, L. (2020). Evaluation of bankruptcy prevention tools: evidences from COSME programme. *Economics Bulletin, 40*(2), 978–988.

Dvouletỳ, O., Srhoj, S., & Pantea, S. (2021). Public SME grants and firm performance in European Union: A systematic review of empirical evidence. *Small Business Economics, 57*(1), 243–263.

European Commission. (2019). *Annual Report on European SMEs 2018/2019* Tech. Rep.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. Academic Press.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11–21.

Lan, M., Tan, C. L., Su, J., & Lu, Y. (2008). Supervised and traditional term weighting methods for automatic text categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(4), 721–735.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

Liberati, C., Camillo, F., & Saporta, G. (2017). Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Advances in Data Analysis and Classification, 11*(1), 121–138.

Liu, B., & Chang, K. C. (2004). Editorial: Special issue on web content mining. *SIGKDD Explor Newsl, 6*(2), 1–4.

Llopis, J., Gonzalez, R., & Gasco, J. (2010). Web pages as a tool for a strategic description of the Spanish largest firms. *Information Processing & Management, 46*(3), 320–330.

Martí, J., & Quas, A. (2018). A beacon in the night: government certification of SMEs towards banks. *Small Business Economics, 50*(2), 397–413.

Mercer, J. (1909). *Functions of positive and negative type and their connection with the theory of integral equations*. London: Philosophical Transactions Royal Society.

Meroño-Cerdan, A. L., & Soto-Acosta, P. (2007). External web content and its influence on organizational performance. *European Journal of Information Systems, 16*(1), 66–80.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B.,& Müller, K.R. (1999). Fisher discriminant analysis with kernels. *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.* (p. 41 -48).

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine, 17*(8), 873–890.

Overbeeke, M., & Snizek, W. E. (2005). Web sites and corporate culture: A research note. *Business & Society, 44*(3), 346–356.

Padilla, P., De Voldere, I., & Duchêne, V. (2018). Is the SME-instrument delivering growth and market creation?. *Assessment of the performance of the first finalized phase II projects*.

Pamukçu, E., Bozdogan, H.,&Çalık, S. (2015). A novel hybrid dimension reduction technique for undersized high dimensional gene expression data sets using information complexity criterion for cancer classification. *Computational and mathematical methods in medicine* 1-14.

Scholkopf, B., Burges, C., & Smola, A. J. (1999). *Advances in Kernel Methods*. MAMIT Press.

Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K. R., Rätsch, G., & Smola, A. J. (1999). Input Space Versus Feature Space in Kernel-Based Methods. *IEEE Transaction on Neural Networks, 5*, 1000–1017.

Séverin, E., & Veganzones, D. (2021). Can earnings management information improve bankruptcy prediction models? *Annals of Operations Research, 306*(1), 247–272.

Suykens, J., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters, 9*(3), 293–300.

Thomaz, C. E., Boardman, J. P., Hill, D. L. G., Hajnal, J. V., Edwards, D. D., Rutherford, M. A., Gillies, D. F., & Rueckert, D. (2004). Using a maximum uncertainty LDA-based approach to classify and analyse MR brain images. Medical Image Computing and Computer Assisted Intervention -MICCAI,. (2004). *Medical image computing and computer assisted intervention - miccai 2004* (pp. 291–300). Berlin: HeidelbergSpringer, Berlin Heidelberg.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer.

Vapnik, V. (1998). *Statistical learning theory*. Wiley.

Wang, L., Jia, F., Chen, L.,& Xu, Q. (2022). Forecasting SMEs' credit risk in supply chain finance with a sampling strategy based on machine learning techniques. *Annals of Operations Research*, 1–33.

Yin, C., Jiang, C., Jain, H. K., & Wang, Z. (2020). Evaluating the credit risk of SMEs using legal judgments. *Decision Support Systems, 136*, 113364.

Zhu, L., & Ou, Y. (2021). Enhance financing for small-and medium-sized suppliers with reverse factoring: a game theoretical analysis.*Annals of Operations Research*, 1–29.