Economics Letters 204 (2021) 109888

Contents lists available at ScienceDirect

Economics Letters

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

Predicting SME's default: Are their websites informative?

Lisa Crosato [a,*], Josep Domenech [b], Caterina Liberati [c]

[a] Department of Economics and Bliss - Digital Impact Lab, Ca' Foscari University of Venice, Cannaregio 873, 30121, Venice, Italy
[b] Department of Economics and Social Sciences, Universitat Politècnica de València, Camí de Vera s/n., 46022 Valencia, Spain
[c] Department of Economics, Management and Statistics (DEMS) and Center for European Studies (CefES), Bicocca University, Milano, Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

ARTICLE INFO

Article history:
Received 14 January 2021
Received in revised form 27 April 2021
Accepted 28 April 2021
Available online 30 April 2021

Keywords:
Default risk
SMEs
Web scraping
Corporate websites
Nonlinear discriminant

ABSTRACT

We propose the use of online indicators, scraped from the firms' websites, to predict default risk for a sample of Spanish firms via nonlinear discriminant analysis and the logistic regression model.

1. Introduction

The Small Business Act of the European Commission in 2008 acknowledges the key role of Small and Medium Enterprises (SMEs) in the European Union economy. In 2018, SMEs constituted 99.8% of all enterprises in the EU-28 non-financial business sector, explaining 56.4% of its value added and 66.6% of its employment (European Commission, 2019).

On the other hand, SMEs experience more difficulties in their early stages mainly due to high market competition and credit constraints (Fritsch and Weyh, 2006). For these reasons, SMEs-addressed support policies are promoted both by the single countries governments and by the European Union (Dvouletỳ et al., 2020; Cultrera et al., 2020).

The literature about SMEs' default prediction covers a variety of case studies and methodologies, starting from the Logistic Regression model (LR) to artificial intelligence and Support Vector Machine models (for a thorough review of studies on bankruptcy prediction see Bellovary et al. (2007)). Most methods were applied on financial and accounting data (Andreeva et al., 2016), while a smaller number of works introduced soft information like the qualitative information on the management (Cornée, 2019) or relational data (Tobback et al., 2017). Other contributions added information on the context in which SMEs operate, such as the local banking market (Arcuri and Levratto, 2020) or the institutional setup (Alexeev and Kim, 2012).

All of these data (offline information) share the drawback of being available with delay with respect to their reference period: accounting indicators, retrieved from balance sheets, are accessible about two years late and qualitative indicators usually derive from personal interviews which require a substantial time amount to be collected, codified and checked. This, in turn, is likely to diminish the promptness of the results, both for credit risk and policy aims, and particularly in a forecasting perspective.

To overcome this issue we propose to predict SMEs' default using data scraped from their corporate websites (online information). This kind of data are available almost in real time, allowing a new approach of monitoring SMEs by automatically detecting their features.

In particular, we work on a sample of companies established in Spain, where SMEs contribution in terms of both employees and value added overcomes the EU-28 average. By way of comparison, we have collected both offline and online information for each firm. The analysis is then carried out first separately on either type of data, and secondly combining both types of data, according to one nonlinear discriminant model benchmarked with the LR model.

Results show that online indicators outperform offline indicators in terms of default prediction, independently on the model.

* Corresponding author.
E-mail addresses: lisa.crosato@unive.it (L. Crosato), jdomenech@upvnet.upv.es (J. Domenech), caterina.liberati@unimib.it (C. Liberati).

https://doi.org/10.1016/j.econlet.2021.109888

2. Data

The sample is composed of 780 Spanish companies, 3.5 percent of which were not active[1] by mid-2016, and thus considered as defaulters. Company information, which refers to the data available in 2014, comes from two sources: (i) offline data, obtained from the balance sheets in the Bureau Van Dijk's SABI (Sistema de Análisis de Balances Ibéricos) database, (ii) online data obtained by web scraping companies websites to automatically extract the most meaningful features, in a process similar to Blazquez and Domenech (2018).

The offline variables selected as predictors are Number of employees, Year of activity, Debt percentage, Productivity and Economic profit, in line with the mainstream literature. With regard to the online variables, Websites were accessed through the Wayback Machine of the Internet Archive which is a digital library of Internet sites able to show the look and the features of a site and its changes over years. Our process summarizes each website as a set of binary indicators representing the presence of a given feature in the website. Indicators include features about the text content of the website and the underlying HTML code (tags or words in the hyperlinks). Text content is related to the public image of the firm, while the HTML code features relate to the technology used and to the up-to-datedness of the website.[2]

The binary indicators were transformed into 41 numerical orthogonal factors, via Multiple Correspondence Analysis (Greenacre, 1984), to allow the application of nonlinear discriminant analysis.

3. Methodology

The overwhelming majority of SMEs credit risk studies used LR (Hosmer and Lemeshow, 2000), also due to its straightforward result interpretation. However nonlinear models generally outperform the LR results (Fantazzini and Figini, 2009; Ciampi and Gordini, 2013; Barboza et al., 2017).

We work with the Kernel Discriminant Analysis (KDA) model, whose mathematical framework converts a nonlinear problem into a linear one by mapping data onto the Feature Space $\mathcal{F}$ (Schölkopf et al., 1999). The latter is defined as the space of all functions mapping from $\mathcal{X} \to \mathbb{R}$, i.e. $\mathbb{R}^{\mathcal{X}} = \mathcal{F} = \{f | f : \mathcal{X} \to \mathbb{R}\}$:

$$\phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}, \mathbf{x} \to k(\cdot, \mathbf{x}) \tag{1}$$

where $\phi(\mathbf{x})$ assigns the value $k(\mathbf{x}', \mathbf{x})$ to $\mathbf{x}' \in \mathcal{X}$, i.e., $\phi(\mathbf{x})(\cdot) = k(\cdot, \mathbf{x})$. We can refer to $\mathcal{F}$ as a Reproducing Kernel Hilbert Space if the Mercer's theorem is satisfied (Mercer, 1909). The *trick* behind Kernel-based methods is to replace dot products in $\mathcal{F}$ with a kernel function in the Input Space, so that the nonlinear mapping is performed implicitly in the new space (Vapnik, 1998). Then, the Fisher's discriminant (Fisher, 1936) is performed in the Feature Space.

Let $\mathcal{I}_{XY} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be the input dataset of training vectors $x_i \in \mathcal{X}$ and the corresponding values of $y_i \in \mathcal{Y} = \{1, 0\}$ the indices characterizing membership to the first and the second class, respectively. The class separability in a direction of the weights $w \in \mathcal{F}$ is obtained maximizing the Rayleigh coefficient:

$$J(w) = \frac{w' \mathbf{S}_B^{\phi} w}{w' \mathbf{S}_W^{\phi} w} \tag{2}$$

[1] This includes firms in: extinction, dissolution, liquidation; in a finished receivership where dissolution or liquidation was ordered, but not yet done; in receivership in progress, except when the firm merged or was taken over.

[2] Further details on the data design and on the websites' indicators can be found in the appendix online.

Table 1
Test set misclassification table: average rates over 100 random samples.

Model		Sensitivity	Specificity	Accuracy
Online data				
RBF	mean *(sd)*	0.706*(0.207)*	0.708*(0.073)*	0.708*(0.056)*
	median *(mad)*	0.778*(0.165)*	0.733*(0.066)*	0.747*(0.052)*
LR	mean *(sd)*	0.183*(0.123)*	0.911*(0.041)*	0.845*(0.038)*
	median *(mad)*	0.167*(0.082)*	0.911*(0.041)*	0.848*(0.030)*
Offline data				
RBF	mean *(sd)*	0.586*(0.185)*	0.757*(0.073)*	0.742*(0.059)*
	median *(mad)*	0.556*(0.165)*	0.789*(0.066)*	0.758*(0.060)*
LR	mean *(sd)*	0.018*(0.041)*	0.987*(0.014)*	0.899*(0.013)*
	median *(mad)*	0.000*(0.000)*	0.989*(0.016)*	0.899*(0.015)*
Online and Offline data				
RBF	mean *(sd)*	0.592*(0.155)*	0.802*(0.052)*	0.783*(0.046)*
	median *(mad)*	0.556*(0.165)*	0.800*(0.066)*	0.788*(0.052)*
LR	mean *(sd)*	0.329*(0.176)*	0.900*(0.045)*	0.848*(0.040)*
	median *(mad)*	0.333*(0.165)*	0.900*(0.049)*	0.854*(0.037)*

where $\mathbf{S}_B^{\phi}$, $\mathbf{S}_W^{\phi}$ are the Between and Within covariance matrices in the $\mathcal{F}$, respectively.

Every solution can be written as an expansion in terms of mapped training data:

$$w = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x}_i) \tag{3}$$

meaning that the Feature Space is accessible via the kernel function only, either because it is too high or infinite dimensional, or it would be impossible to obtain an explicit solution $w \in \mathcal{F}$ (Mika et al., 1999). We apply the widely used Radial-Basis Function (RBF) defined as $\phi(\mathbf{x}_i) = (-\exp \|\mathbf{x}_i - \mathbf{x}\|^2/\sigma)$ for some $\sigma \geq 0$, with grid search parameter optimization.

4. Results

Our classification protocol splits the data into training (220 instances) and test set (99 instances), which are randomly selected from the original sample. To get robust classification rates, we bootstrapped the dataset 100 times and collated average rates of prediction of default (sensitivity), survival (specificity) and total correct classification (accuracy), according to both the mean and the median and the corresponding measures of variability (Table 1). Survived firms were undersampled, in the ratio of ten survivors to one defaulter, to reduce the unbalance between the two groups.

Since the focus of the paper is on predicting default, we are particularly interested in the sensitivity figures.

The RBF kernel applied to online indicators allows for a correct prediction of default in 70.6% of the cases, rising to 77.8% when considering the median sensitivity. Online indicators outperform offline indicators according to both measures. The loss in specificity and, in turn, in accuracy, when moving from offline to online indicators is not as large as to balance the gain in correctly classifying the defaulters.

On the contrary, the default classification resulting from the LR model applied to both kind of data is unsatisfactory due to the unbalance between the two groups (although reduced by our sampling), which affects performances of the LR. In fact the accuracy reaches 90% with the offline indicators and about 85% with the online indicators) but the sensitivity is under 2% (mean) and is 0 according to the median for offline indicators. Note however that the same model applied to online indicators increases the

mean sensitivity up to 18.3% (16.7% median). Therefore, using online indicators reduces the disequilibrium between classification of survivors and defaulters, regardless the method.

The rate variability is similar between online and offline data, particularly when considering the mad, and always smaller in accuracy with online data.

Combining online and offline indicators increases specificity but keeps sensitivity under 60% according to RBF, while it enhances sensitivity by the LR model.

5. Conclusions

This short study provides some encouraging evidence for the use of online indicators retrieved from the firms' websites to predict businesses' default. Overall, present results are not in favor of the joint use of online and offline data: studying alternative combinations of both types of indicators to exploit possible common pieces of information calls for further research (see the Supplementary Data Appendix).

An extension of the same framework of analysis to more kernel functions, large datasets and, possibly, to several countries is also on the agenda.

Assessing borrowers creditworthiness in the banking sector is the natural application for this methodology. The inclusion of the online predictors in a dynamic setting could be tested in a nowcasting perspective as corporate websites can be accessed anytime to gather information on the appearance of new products and services, updates in the technology or in the communication with customers. Online indicators could also complement or substitute accounting figures when the latter imply less accurate default prediction for SMEs with respect to larger firms (see Ciampi and Gordini (2013)), or in credit evaluation of microfirms for which accounting information is limited and the use of soft information prevails (Altman et al., 2010).

All of the above cases apply not only to bank lending practice but also to government policies for SMEs supporting.

Acknowledgments

This work was partially supported by the Ca' Foscari University of Venice, Italy and by Agencia Estatal de Investigación, Spain under grant PID2019-107765RB-I00. We also acknowledge helpful comments by an anonymous referee.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.econlet.2021.109888.

References

Alexeev, M., Kim, J., 2012. Bankruptcy and institutions. Econ. Lett. 117 (3), 676–678.
Altman, E.I., Sabato, G., Wilson, N., 2010. The value of non-financial information in small and medium-sized enterprise risk management. J. Credit Risk 6 (2), 1–33.
Andreeva, G., Calabrese, R., Osmetti, S.A., 2016. A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. European J. Oper. Res. 249 (2), 506–516.
Arcuri, G., Levratto, N., 2020. Early stage SME bankruptcy: does the local banking market matter? Small Bus. Econ. 54 (2), 421–436.
Barboza, F., Kimura, H., Altman, E., 2017. Machine learning models and bankruptcy prediction. Expert Syst. Appl. 83, 405–417.
Bellovary, J.L., Giacomino, D.E., Akers, M.D., 2007. A review of bankruptcy prediction studies: 1930 to present. J. Financ. Educ. 33, 1–42.
Blazquez, D., Domenech, J., 2018. Web data mining for monitoring business export orientation. Technol. Econ. Dev. Econ. 24 (2), 406–428.
Ciampi, F., Gordini, N., 2013. Small enterprise default prediction modeling through artificial neural networks: An empirical analysis of Italian small enterprises. J. Small Bus. Manag. 51 (1), 23–45.
Cornée, S., 2019. The relevance of soft information for predicting small business credit default: Evidence from a social bank. J. Small Bus. Manag. 57 (3), 699–719.
Cultrera, L., et al., 2020. Evaluation of bankruptcy prevention tools: evidences from COSME programme. Econ. Bull. 40 (2), 978–988.
Dvouletỳ, O., Srhoj, S., Pantea, S., 2020. Public SME grants and firm performance in European Union: A systematic review of empirical evidence. Small Bus. Econ. 1–21.
European Commission, 2019. Annual Report on European SMEs 2018/2019. Technical Report.
Fantazzini, D., Figini, S., 2009. Default forecasting for small-medium enterprises: Does heterogeneity matter? Int. J. Risk. Assess. Manag. 11 (1–2), 138–163.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7 (2), 179–188.
Fritsch, M., Weyh, A., 2006. How large are the direct employment effects of new businesses? An empirical investigation for West Germany. Small Bus. Econ. 27 (2), 245–260.
Greenacre, M.J., 1984. Theory and Applications of Correspondence Analysis. London (UK) Academic Press.
Hosmer, D., Lemeshow, S., 2000. Applied Logistic Regression. Wiley.
Mercer, J., 1909. Functions of positive and negative type, and their connection the theory of integral equations. Philos. Trans. R. Soc. Lond. Ser. A 209, 415–446.
Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Muüller, K.-R., 1999. Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp. 41–48.
Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Muüller, K.-R., Rätsch, G., Smola, A.J., 1999. Input space versus feature space in kernel-based methods. IEEE Trans. Neural Netw. 5, 1000–1017.
Tobback, E., Bellotti, T., Moeyersoms, J., Stankova, M., Martens, D., 2017. Bankruptcy prediction for SMEs using relational data. Decis. Support Syst. 102, 69–81.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley.