4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15103

Non-conventional data and default prediction: the challenge of companies' websites

Lisa Crosato[1], Josep Domenech[2], Caterina Liberati[3]
[1]Department of Economics, Ca' Foscari University of Venice, Italy, [2]Department of Economics, Universitat Politècnica de València, Spain, [3]Department of Economics Management and Statistics, University of Milano-Bicocca, Italy.

Abstract

Small and Medium Enterprises (SMEs) contribution to the European Union economy has always been relevant, for both value added and the creation of jobs. That is why the prediction of their survival is considered one of the economic pillars UE keeps under observation. Default prediction models, accounting for SMEs idiosyncratic traits, are based on several types of data, mainly accounting indicators. Balance sheet data, indeed, are considered the standard predictors for classification models in this field, although they do not allow to completely overcome the information opacity that is one of the main barriers preventing these firms from accessing credit. In our work, we explore the possibility of complementing accounting information with data scraped from the firms' websites. We modeled the data using a nonlinear discriminant analysis and we benchmarked the results with the Logistic Regression. The evidence of our study is promising although the combination of online and offline data shows better results in case of survival firms than for defaulted companies.

Keywords: *Website Data; SMEs; Default Prediction; Kernel Discriminant.*

1. Introduction

Economic studies on businesses present a wide literature on forecasting SMEs default. The great interest that scholars and practitioners have been showing toward this particular topic is given by the combination of two aspects: first, SMEs are 99% of all enterprises in the European Union, covering the largest part of the European value added and jobs (56.4% and 66.6% respectively, European Commission, 2019); second, the access to credit for these companies is difficult, especially in their early stages, due to information opacity. This makes the assessment of the creditworthiness of a SME a relevant issue, particularly in a policy maker and credit lender perspective (Cultrera, 2020; Belghitar et al, 2021).

To get low misclassifications' rates between survival and defaulted companies, researchers has focused mainly on the derivation of credit-scores based on Machine Learning (ML) algorithms (eg. Random Forest, Support Vector Machines), because they have shown the best performances with respect to the standard linear classifiers (Baesens et al, 2003; Lessmann et al, 2015). Input variables of such models are, generally, accounting indicators derived from balance sheets (Fantazzini & Figini, 2009; Succurro & Mannarino, 2014; Ciampi, 2015) provided by databases as Bureau van Dijk. Indeed, the quality of these data are really high - thanks to well established data-collection procedures- but on the same time, they suffer of a large delay between their availability and their reference period. Unfortunately, this drawback prevents a prompt prediction of default, diminishing the value of the results in a forecasting perspective.

In our paper we propose the use of websites as an additional source of information for detection of SMEs default (Crosato et al, 2021). The analysis of corporate websites to get new proxies of the standard business indicators has been already investigated in previous works: for capturing different corporate culture dimensions (Overbeeke & Snizek, 2005), or for measuring firm performances (Merono-Cerdan & Soto-Acosta, 2007) and the level of innovation (Axenbeck & Breithaupt, 2021).

The employment of these unconventional data requires an articulated architecture of data pre-processing, including specific procedures for data retrieval, cleaning and dimension reduction. Naturally, the knowledge extraction process is not straightforward: it is directly correlated with the complexity of the analyzed data. On the other hand, the gain in terms of additional information, which is available and free, largest coverage of the firms' population as well as the recency of the obtaining data rewards the analytical efforts.

The web-based indicators could be generated reviewing the corporate websites one by one using a manual evaluation of the features (Blazquez & Domenech, 2018), although it is very time consuming and not recommended when monitoring a high number of companies. In this study we apply instead an automatic process that photographs the websites at a given time and also tracks their changes. Using the Wayback machine, a digital archive of the World

Wide Web, we were able to see archived versions of web pages across time and compare their evolutions.

We analyze a sample of 700 SMEs sampled from the SABI - Sistema de Análisis de Balances Ibéricos (Bureau van Dijk). The database we built merges the accounting (offline) indicators and the web-based (online) indicators. We aim to verify whether website features help to predict corporate bankruptcy and in particular which indicators, among the online ones, can be selected to discriminate between surviving and defaulted firms. We also study if the joint use of online and offline information aids to reduce misclassifications error rates both using a nonlinear discriminant model and a Logistic regression.

2. Websites Data

To understand the process to create online indicators, it is necessary to illustrate how websites are built and organized. A website is a set of documents stored in a web server. The Hyper Text Markup Language (HTML) is the language used for setting the formatting and the layout of the hypertextual documents, including all the specifications of the webpages structure. A corporate website can be studied using two approaches: mining the web structure or the web content. The two approaches focus on different characteristics of the sites: the first one detects the linkages among the web-pages, the second one concentrates on understanding the semantics and the meaning of the contents. In our paper we rely on the latter, because it best describes the business activity.

In practice, we carry out the study from a twofold perspective:

- **Textual analysis** - It is useful to retrieve all the relevant information shared on the website by the companies: the economic sector in which they operate, market orientation (e.g., national or international, final consumer or other businesses) the locations and the activities done. Additionally, any editing/updating in the text can be reviewed as a sign of investment in the online communication and consequently of active behavior. The textual analytics, which encompasses a large variety of data manipulation processes (eg. cleaning and words stemming), helps to identify meaningful terms with the highest occurrences. This way, the unstructured text is converted into a set of dummies variables that can be later used by the classifiers.

- **HTML code**- The analysis of the HTML code provides important information about the tag of a webpage. The tags carry knowledge about the interaction (e.g., defining hyperlinks or forms), appearance (e.g., bold or italics), and the structure (e.g., defining lists or different blocks) of a web page. They evolve according to the progress of HTML language, so they indirectly represent the complexity and the level of technology of a

website. For instance, EMBED is generally employed to include Flash technology, which is currently being abandoned; FORM is usually employed to interact with the company/site. The hyperlinks connections are listed as tag A, they include all the internal/external connections present in the website (href), the extensions of the files shared (pdf, excel, word) and the underlying technology (php/asp/htm). The images are listed as tag IMG with their correspondent extension.

3. Online features

Before dealing with any classification model, we start by simply reducing the number of available features to the ones showing significant differences between the groups of survived and the group of defaulted SMEs. To this purpose we calculate, within each group, the proportion of firms on whose website the considered feature was present and then apply a standard test for the difference in proportions to each of the features. Significant differences were found for three main groups of features: 7 in the category "Hrefwords", 20 in the category "Stems" and 18 in the category "Words", plus 4 in a residual miscellaneous category. The majority of features making the difference appear more often in the survived group, particularly in the "Hrefwords" category.

4. Methods

The online indicators we are working with are now binary and this could restrict the application of certain models. Therefore, we transform them into numerical orthogonal factors via Multiple Correspondence Analysis (Greenacre, 1984). As about accounting indicators, the literature suggests to take care of possible associated non-linear patterns, and we expect that including the online features in the analysis will add complexity to the within-variable relationships. Another aspect compounding the classification task is the overbalance between survived and defaulted companies. For assessing both issues, we refer to the wide range of statistical techniques in the relevant literature on SMEs bankruptcy prediction.

Machine-learning methods have been generally found to outperform the linear ones. Non-linear models such as Deep Learning (Mai et al., 2019), Boosting (Kou et al., 2021) and Neural Networks (Baesens et al., 2003) have been successfully employed, maintaining the z-score (Altman, 1968) or the logistic regression (Hosmer and Lemeshow, 2000) as a benchmark. Here we use Kernel Discriminant Analysis (KDA, Mika et al., 1999) due to very good performances of kernel-based algorithms in screening SMEs (Gordini, 2014; Zhang, 2015).

The goal of KDA is to provide a decision function *f(x)* from the combination of features that best separates two classes of objects. Given a training set $I_{XY} = \{(x_1, y_1), \dots, (x_n, y_n)\}$

describing *n* units, with data $x_i \in R_p$ and labels $y_i \in \{0,1\}$, due to the recurrent non-linear separability of training data in the input space R_p, the training data are usually associated to some feature space *F* via a non-linear mapping:

$$\varphi: x_i \rightarrow \varphi(x_i) \qquad (1)$$

F can be referred to as a Reproducing Kernel Hilbert Space (RKHS) when the Mercer's theorem is satisfied (Mercer, 1909).

The ratio of the Between and Within covariance matrices (computed in the Feature Space) is then minimized to obtain a separating hyperplane in *F*, as in the Fisher Discriminant Analysis (Fisher, 1936):

$$g(x) = w^T \varphi(x) + b \qquad (2)$$

where *w* is the weight vector in RKHS, and $b \in R$ is the bias term.

Getting an optimal generalization of kernel-based methods still requires choosing a suitable kernel map. Among the many proposed in the literature we have selected the Radial Basis Function (RBF), Laplace, Cauchy and Multiquadric kernels due to their remarkable performances.

Acknowledgments

This work was partially supported by grants PID2019-107765RB-I00 and funded by MCIN/AEI/10.13039/501100011033.

References

Altman, E.I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance 23(4), 589–609

Axenbeck, J. & Breithaupt, P. (2021). Innovation indicators based on firm websites - which website characteristics predict firm-level innovation activity? PloS one 16(4), e0249, 583

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(6), 627–635

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 pp. 1–19

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

Ciampi, F. (2015) Corporate governance characteristics and default prediction modeling for small enterprises. an empirical analysis of Italian firms. Journal of Business Research 68(5), 1012–1025

Crosato, L., Domenech, J. & Liberati, C. (2021) Predicting SMEs default: Are their websites informative? Economics Letters 204, 109,888

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

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

Fantazzini, D. & Figini, S. (2009) Default forecasting for Small-Medium Enterprises: Does heterogeneity matter? International Journal of Risk Assessment and Management 11(1-2), 138–163

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. Ann. Eugen. 7(2), 179–188

Gordini, N. (2014)A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Systems with Applications 41(14), 6433–6445

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

Hosmer, D., Lemeshow, S. (2000) Applied Logistic Regression. Wiley

Kou, G., Xu, Y., Peng, Y., Shen, F., Chen, Y., Chang, K., Kou, S. (2021) Bankruptcy prediction for SMEs using transactional data and two-stage multiobjective feature selection. Decision Support Systems 140, 113,429

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

Mai, F., Tian, S., Lee, C., Ma, L. (2019) Deep learning models for bankruptcy prediction using textual disclosures. European Journal of Operational Research 274(2), 743–758.

Mercer, J. (1909) Functions of positive and negative type, and their connection the theory of integral equations. Philosophical transactions of the royal society of London, A 209, 415–446.

Mika, S., R*ä*tsch, G., Weston, J., Sch*ö*lkopf, B., M*ü*ller, 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

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

Succurro, M. & Mannarino, L. (2014). The Impact of Financial Structure on Firms' Probability of Bankruptcy: A Comparison across Western Europe Convergence Regions. Regional and Sectoral Economic Studies, Euro-American Association of Economic Development, 14(1), 81–94

Zhang, L., Hu, H., Zhang, D. (2015) A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance. Financial Innovation 1(1), 1–21