4th International Conference on Advanced Research Methods and Analytics (CARMA2022)
Universitat Politècnica de València, València, 2022
DOI: http://dx.doi.org/10.4995/CARMA2022.2022.15045

Can unlisted firms benefit from market information? A data-driven approach

Alessandro Bitetto[1], Stefano Filomeni[2], Michele Modina[3]
[1]Department of Economics and Management, University of Pavia, Italy, [2]Essex Business School, Finance Group, University of Essex, United Kingdom, [3]Department of Economics and Management, University of Molise, Italy.

Abstract

We employ a sample of 10,136 Italian micro-, small-, and mid-sized enterprises (MSMEs) that borrow from 113 cooperative banks to examine whether market pricing of public firms adds additional information to accounting measures in predicting default of private firms. Specifically, we first match the asset prices of listed firms following a data-driven clustering by means of Neural Networks Autoencoder so to evaluate the firm-wise probability of default (PD) of MSMEs. Then, we adopt three statistical techniques, namely linear models, multivariate adaptive regression spline, and random forest to assess the performance of the models and to explain the relevance of each predictor. Our results provide novel evidence that market information represents a crucial indicator in predicting corporate default of unlisted firms. Indeed, we show a significant improvement of the model performance, both on class-specific (F1-score for defaulted class) and overall metrics (AUC) when using market information in credit risk assessment, in addition to accounting information. Moreover, by taking advantage of global and local variable importance technique we prove that the increase in performance is effectively attributable to market information, highlighting its relevant effect in predicting corporate default.

Keywords: *credit risk, distance to default, machine Learning, market information, probability of default.*

1. Introduction

The aim of banks' core business is to perform accurate assessment of borrowers' capability to repay their debt by collecting information about a given borrower from different sources. The type of information a bank should use when assessing credit risk has been a matter of concern for policy makers since inaccurate credit risk measurement could threaten the stability of the banking sector. In this regard, banks' need to implement reliable credit risk models to timely and precisely forecast business failure is imperative to reach appropriate lending decisions and, eventually, to engage in corrective action.

When focusing on the predictions of default risk of micro-, small- and mid-sized enterprises (MSMEs), a credit risk assessment model should take into account their peculiarities which are not similar to those of larger firms. MSMEs exhibit higher default risk and greater information opacity. Given their importance for market economies, it is imperative to implement credit assessment models specifically addressed to MSMEs with the objective to minimize expected and unexpected losses as accurately as possible.

In this paper, we develop a credit risk model for MSMEs that considers, in addition to accounting measures, market information obtained from comparable publicly listed companies adopting three statistical techniques, namely linear models, multivariate adaptive regression spline, and random forest. Assembling a comprehensive dataset that includes 10,136 unlisted Italian MSMEs, we estimate multivariate forecasting models on the incidence of corporate default by using both market and accounting information employing several advanced statistical techniques. Given the nature of our dataset, we estimate the Merton's Probability of Default (PD) based on market information obtained from listed companies and deemed as comparable by a data-driven clustering approach, avoiding any a-priori assumption of mapping by size, industry and number of employees.

The paper contributes to the literature along two dimensions. The first one involves the implementation of predictive models and their explainability. Our work contributes to a new stream of research (usually called eXplainable Artificial Intelligence) by implementing both a non-linear parametric and non-parametric ML algorithms. Specifically, we go beyond the forecasting of corporate defaults and implement an advanced methodology that involves the use of two cutting-edge techniques to evaluate the importance of variables on forecasts: Permutation Feature Importance (Fisher et al., 2018) explains the overall variables' relevance, whereas Shapley Additive Explanations (Lundberg et al., 2020) provide the contribution of each variable's values to the predicted probability of default for a single observations. In addition, we implement a sophisticated clustering technique that, to the best of our knowledge, is the first application of Artificial Neural Networks to compress the information of financial ratios so to map each unlisted MSMEs to a pool of listed ones. Secondly, our hybrid credit scoring models, which use a combination of market

and accounting information, provide better default predictions for unlisted firms when compared with the respective predictive power of models which only use accounting or market information. We demonstrate that the estimated Merton default probability (PD) measure has incremental predictive power over corporate default when added to a multivariate predictive regression model that already includes accounting information.

One policy implication resulting from our findings is that banks can potentially integrate their hybrid credit scoring methodologies with market information for credit risk assessments, with the purpose of increasing the accuracy of forecasting corporate defaults for unlisted firms. This would allow banks to expand the spectrum of information used in credit risk measurements helping them to enhance their internal hybrid credit scoring by including both accounting and market information on the credit quality of a given borrower. Thus, results reported in this paper could be very helpful for forward-looking financial risk management frameworks (Rodriguez Gonzalez et al., 2018).

The remainder of this paper is organized as follows. Section 2 discusses the data and Section 3 presents the econometric methodology. Section 4 illustrates the empirical results.

2. Data

We use two sources of information for our analysis: a proprietary one, consisting of granular information of 10,136 Italian unlisted MSMEs, and a public one, comprising data on comparable publicly listed companies, i.e., peers.

2.1. MSMEs data

We exploit a unique and disaggregated dataset on an unbalanced panel sample of 10,136 firms and 113 cooperative credit banks, for a total of 19,743 firm-year observations over the period 2012–2014. Specifically, we consider firms with less than 250 employees and revenue at most of 50 million. We selected a subset of 22 financial ratios out of 30 removing the ones showing high partial correlation with many other ratios. Therefore, some ratios with mild correlation with at most one other ratio are still kept because the models we use for the predictions are robust to multicollinearity.

2.2. Peers Data

We select a panel of 40 Italian listed firms, evenly distributed in manufacturing and services sector. We collect accounting figures from Orbis database, developed by Bureau Van Dijk (a Moody's analytics company), by matching the VAT code for each given peer firm . The accounting figures are used to reconstruct and match or proxy the 22 financial ratios of the MSMEs dataset. Moreover, daily stock prices are collected from Refinitiv

Eikon database and are used to compute the annual assets volatility of comparable publicly-listed companies.

3. Methodology

The aim of this paper is to assess the impact of market information, i.e., the Merton's probability of default (PD), in predicting corporate default risk of unlisted firms, in addition to accounting based measures. Our analysis can be summarized into three steps. Firstly, we match each MSME to one or a group of peers and evaluate its firm-wise PD. Section 3.1 recalls how the PD is evaluated following the Merton's model and Section 3.2 describes the peers-to-firm matching procedure, consisting of a low dimensional representation of the 22 variables space and its subsequent clustering. Secondly, we predict corporate default by calibrating different classification models, both using financial ratios as predictors (baseline) and including the PD (extended). Section 3.3 shows the calibration of the models and the differences of models' performance between the baseline and extended cases. Lastly, we investigate which predictor contributes the most to predict corporate default, by means of feature importance techniques. Section 3.4 reports the estimation of the contribution of each variable to the predicted class (default or non-default) for both the baseline and extended cases.

3.1. Estimation of the Merton model

We estimate the Merton model of corporate default risk for our sample of MSMEs. According to the Merton model, the corporate default takes place when the company is unable to pay off its debts, or when the current market of assets falls below the market value of liabilities. For this reason, the market value of equity of the MSME is treated as a call option on the asset value of the MSME with strike price equal to the market value of debt . The MSME asset value process follows a Geometric Brownian motion as shown in Equation (1) below:

$$dA_t = rA_t dt + \sigma_A A_t dz \tag{1}$$

where At is the firms market value of assets and σ_A is the volatility of assets, r is one-year maturity risk-free rate of return, which we choose to be the yield of the 1-year maturity domestic government bond with 1-year maturity .

3.2. Matching unlisted firms with peers

Since there are no market data available for our sample of unlisted MSMEs, we proxy the market volatility of assets of unlisted MSMEs with those of their comparable publicly-listed companies. As for the latter, the market value of assets is computed as the daily

product of their share price multiplied by the number of shares outstanding. Our implicit assumption made for the estimation of the Merton's Probability of Default (PD) and Distance-to-Default (DD) is that those MSMEs which operate in the same industry sectors and have similar balance sheet behaviour with our Italian peers share the same risk profile and belong to the same (market) risk class of the latter . In order to render the matching procedure as accurate as possible, we opt for a clustering approach: we find the optimal number of clusters in the MSME dataset and then we assign each peer to the most similar cluster by minimizing the average distance from all firms in the cluster.

3.3. Prediction of default

After assigning the PD to all our unlisted MSMEs, we calibrate three different models to predict the binary target, (1) for defaulted firm and (0) otherwise. Each model is calibrated with the set of 22 variables (baseline) and with the addition of the PD (extended). First, we inspect the distribution of each input variable with respect to the target variable. Second, we opt for a non-linear and piecewise model, the Multivariate Adaptive Regression Spline (MARS), that estimates multiple polynomial relationships in different partition intervals of each input variable. So, the model can be seen as an ensemble of sub-models that are estimated in each combination of partitions in which input variables can be divided. As MARS is a parametric algorithm, meaning that we have to define a structure of each estimation function, e.g. polynomial, we test also a non-parametric model, the Random Forest (RF).

3.4. Importance of variables

We explore which input variable contributes the most in each model predictions, focusing on the changes when the PD is added. For this reason, we evaluate the predictive power of the variables using two state-of-the-art techniques for feature importance: Permutation Feature Importance (PFI) and Shapley Additive Explanations (SHAP). PFI evaluates the importance of the j-th variable by comparing the performance, e.g. F1-score, of the model that predicts the observations used for the calibration against the performance of the model that predicts the same observations where the values of the j-th column are shuffled. In this way the correlation between the j-th variable and all the others is broken thus removing the influence of that variable on the model predictions. If the change in performance is negligible, the j-th variable is not important for the model. SHAP is based on Shapley values, a method from coalitional game theory which provides a way to fairly distribute the payout among the players by computing average marginal contribution of each player across all possible coalitions. SHAP, uses Shapley values to evaluate the difference of the predicted value of a single observation, comparing the prediction of all possible combinations of variables that include the j-th variable against the ones that do not. The differences are then averaged and the positive or negative change in the prediction is used

as variable importance. For example, if the model predicts the probability of default, SHAP evaluates, for a single observation, which variable contributed most to increase/decrease the final probability. In this way, by exploiting the additive property of Shapley values, it is possible to estimate the impact of all variables on the final predicted value, for each single observation. PFI provides a global measure of importance by assessing the impact of all observations together. Moreover, it measures the changes of a global performance. SHAP, on the other way, provides a local measure of importance, measuring the impact of variables for every single observation. However, taking the average of the absolute values of each observation's SHAP, it is still possible the get a global measure of the average importance of the variables. Instead, taking the average of the Shapley values rather than their absolute value, provides an average effect of each variable on the predictions.

4. Results

As described in Section 3.2, we firstly find the embedding that minimizes the Reconstruction Error. Table 1 reports the optimal embedding dimension *k*, the reconstruction error of the different algorithms and the R^2. In our context, in analogy with the classical R^2, we compute the RSS term as the Reconstruction Error given by the embedding and the TSS term as the total variance contained in the original data and represents a proxy of how much intrinsic information within the data is preserved in the transformation.

Table 1. Results of dimensionality reduction

Input level	Rows	Columns	Method	Input Dimension	Embedding Dimension	Reconstruction Error (% of Avg Abs Input)	R^2
Firm-year	Firm-year pairs	Variables	AE	19,743 x 22	19,743 x 6	0.1418 (20%)	98%
			RobPCA	19,743 x 22	19,743 x 9	0.2033 (30.6%)	95.70%
Firm (batch of years)	Firms	Variables	AE-LSTM	10,136 x 22	10,136 x 10	0.2138 (31.8%)	94.60%
Firm	Firms	Variables-year pairs	AE	10,136 x 66	10,136 x 32	0.2391 (35.9%)	91.30%
			RobPCA	10,136 x 66	10,136 x 15	0.3857 (58%)	84.80%

Source: our elaboration.

The embedding resulting from AE (AutoEncoder) with the firm-year level approach performed best showing the lowest reconstruction error and the highest R^2. Methods evaluated with firm level approach performed worst and won't be included in the following analysis. Then, we look for the optimal number *C* of clusters. We select $C = 5$ clusters identified on the AE embedding. Moreover, we apply the UMAP algorithm to visualize the

clusters into a 3-dimensional space. Figure 1 depicts the five clusters for all observations (small points) as well as the matched peers (bold spheres), showing a good separation, even if there is small overlapping between the yellow and green cluster and few blue peers are mapped close to the red ones. We recall that the embedding function *f* is estimated only on the MSMEs dataset and then the peers' embedding is evaluated by applying *f*. Being the PD assigned, we calibrate the prediction models. The following results refer to the PDs evaluated with the pointwise-PD approach because it performed better than the average-PD one, although the findings described below still hold robust. We tune the parameters of each model with the Stratified Cross-Validation and we calibrate the models with the optimal parameters on the entire dataset, so to have a single model to be used for feature importance evaluation. In Table 2 we report the performance on the entire dataset and the average performance on validation folds for each model as well a comparison between the models trained with the 22 ratios only and the ones with the addition of PD. Random Forest is the only model with good performances, being able to capture the different local separation of the data, as discussed in Section 4.3. Nevertheless, all models show an improvement on class-specific performance, i.e. F1-score for the defaulted class, and on the AUC when the PD is included as predictor.

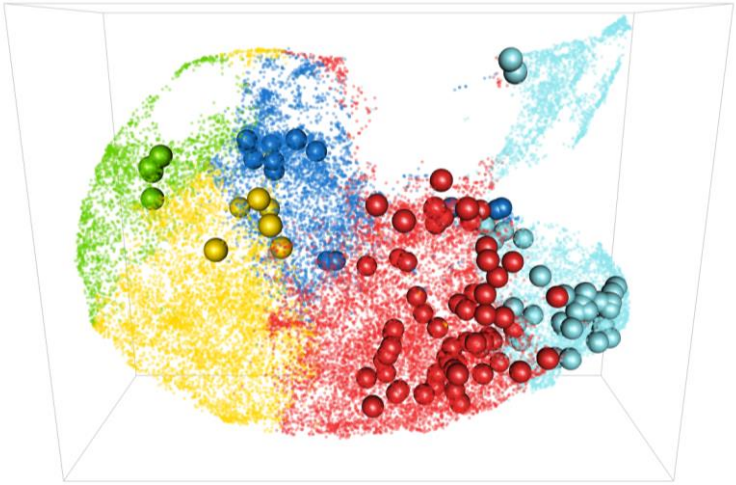

Figure 1. 3D visualization of five clusters for the 6-dimensional AE embedding. Source:our elaboration.

Finally, we explore the feature importance for all models. PFI and SHAP are evaluated on model calibrated with input variables and with the addition of PD. Figure 2 shows the PFI of Random Forest model, where the changes of F1-score are normalized. PD is the second most important variable, slightly below the financial interest on revenues.

Table 2. F1-score and AUC for Elastic-Net, MARS and Random Forest calibrated on dataset with input variables only and with the addition of PD

	F1 (Cross-Val)		AUC (Cross-Val)	
Algorithm	Baseline	With PD	Baseline	With PD
Elastic-Net	30.7% (30.1±1.7%)	35.1% (35.1±1.5%)	79.8% (79.6±0.6%)	82% (81.7±0.8%)
MARS	36% (33.8±1.4%)	40% (37.5±0.6%)	82.5% (81.7±0.6%)	84.2% (82.8±0.8%)
Random Forest	89.5% (85.1±1.7%)	95.8% (91.4±1.2%)	89.8% (85.4±1.1%)	96.1% (91.7±0.7%)

Source: our elaboration.

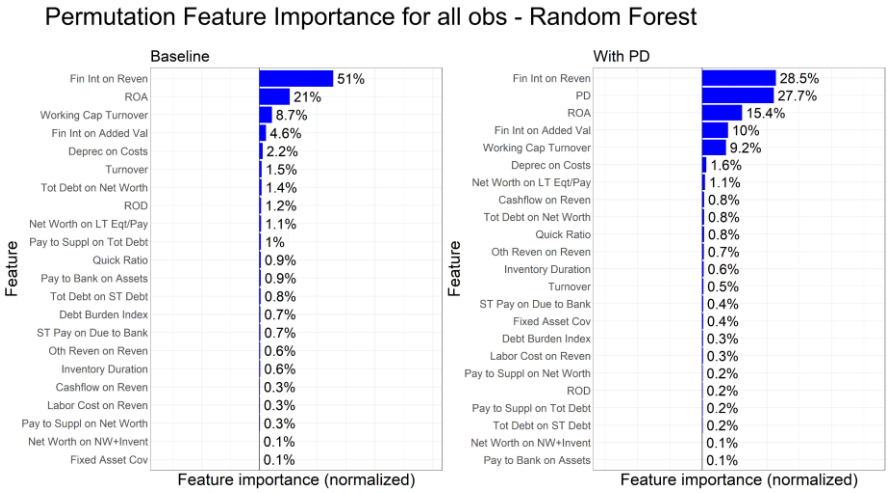

Figure 2. Permutation Feature Importance for Random Forest model. Source:our elaboration.

References

Fisher, A., Rudin, C., and Dominici, F., 2018. odel class reliance: Variable importance measures for any machine learning model class, from the 'rashomon' perspective. URL http: //arxiv.org/abs/1801.01489.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I., 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:2522–5839.

Rodriguez Gonzalez, M., Basse, T., and Kunze, F., 2018. Early warning indicator systems for real estate investments: Empirical evidence and some thoughts from the perspective of financial risk management. *ZVersWiss*, 107:387–403.