Machine Learning for Zombie Hunting: Predicting Distress from Firms' Accounts and Missing Values

Madrid, November 24th, 2023

Prof. Massimo Riccaboni

IMT - School for Advanced Studies
Lucca, Italy

Missing values in economics: Why should I care?

- Missing values are usually attributed to human error when processing data, respondents' refusal to answer certain questions, drop-out in studies, and merging unrelated data
- The seriousness of missing values depends in part on (1) **how much** data is missing, (2) **the pattern** of missing data, and (3) **the mechanism** underlying the missingness of the data
- Missing values can be handled by certain techniques including, deletion of instances, replacement with potential or estimated values, using missingness in attributes
- 95% of economists typically opt for a complete case (CC) analysis, and 4,9% consider imputation techniques, **this presentation is about the rest**.

Missingess not at random

- Dealing with missing data is crucial since improper handling may lead to drawing biased inferences (Little & Rubin, 2019)
- Three types of missingness patterns:
  1. **Missing completely at random (MCAR)**. Missing values are independent both of observable variables and of unobservable parameters of interest. CC is unbiased (it rarely happens).
  2. **Missing at random (MAR)**. Missingness can be fully accounted for by variables where there is complete information: small firms are less likely to provide financial information but this has nothing to do with their chance to survive after accounting for their size (is it reasonable?)
  3. **Missing not at random (MNAR)**. The value of the variable that's missing is related to the reason it's missing: firms are less likely to disclose critical information.
- MNAR is very common and the CC approach popular in economics leads to biased estimates!

Missingness in causal panel models

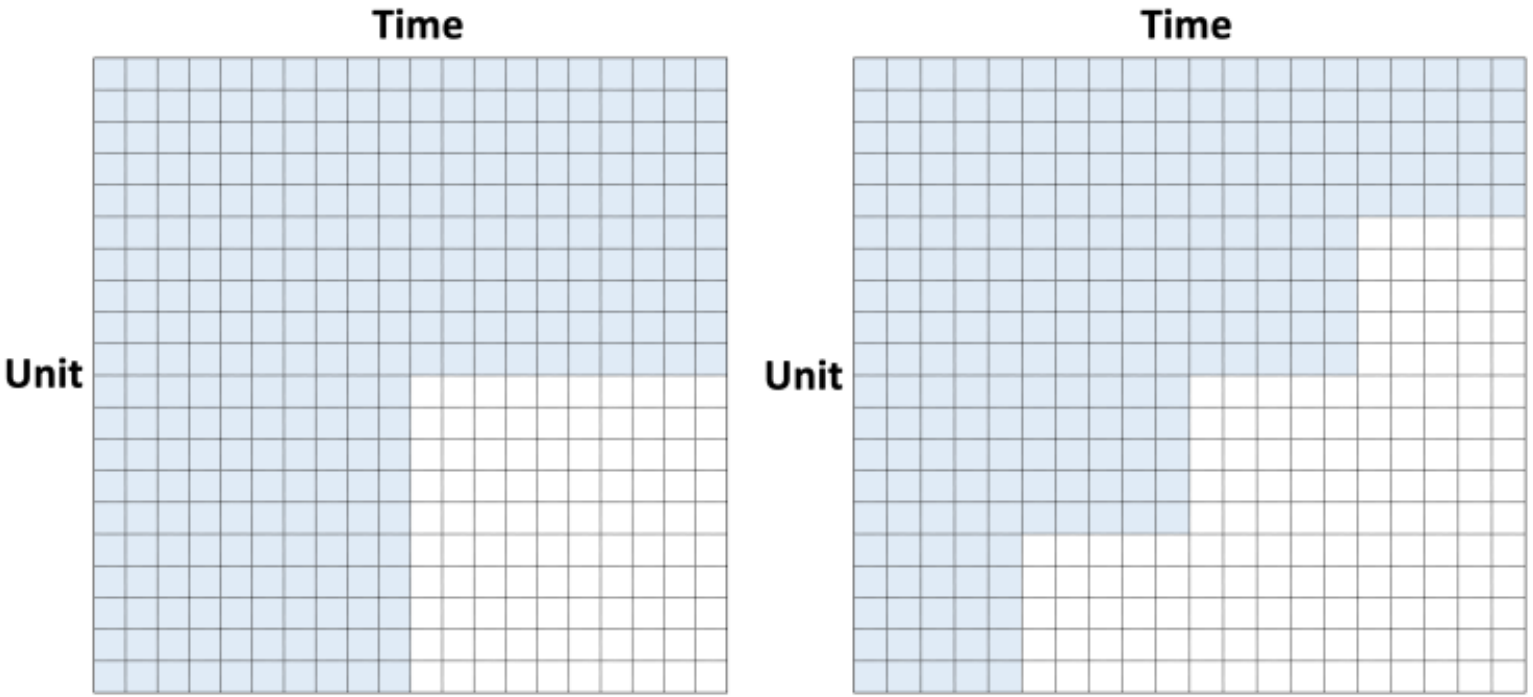
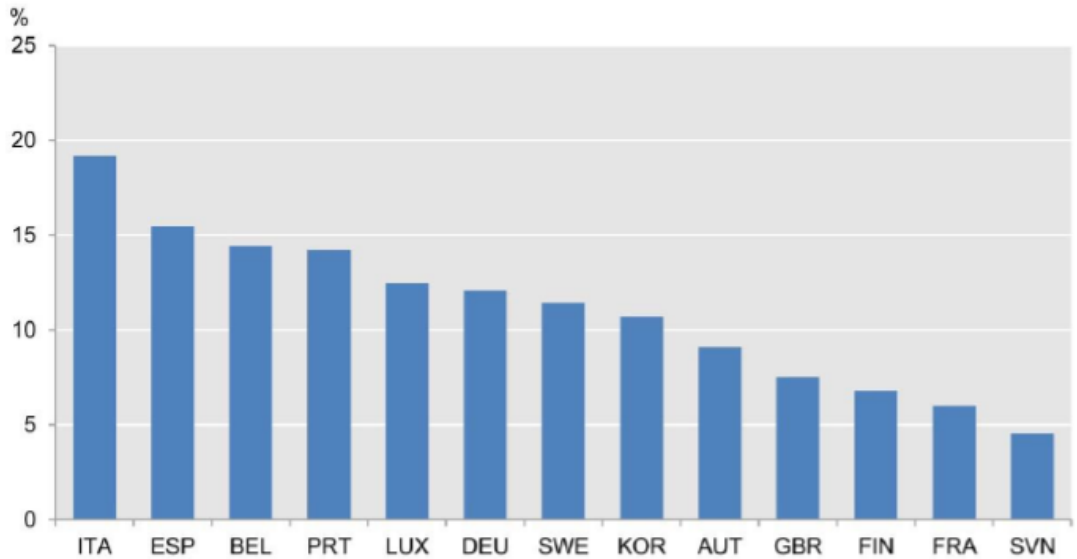


Figure: Two typical observation patterns of the potential outcomes under the control in the causal panel model: Here, the blue area is the observed area, and the white area is the missing area. Missingness occurs because we cannot observe the potential outcomes under the control of the treated entries.

Zombie firms: why should I care?

- Since the global financial crisis and again after the Covid crisis many countries have struggled with economic recovery despite unprecedented stimuli by central banks and national governments.
- A major challenge in recoveries are **zombie firms**.



Note: Firms aged ≥10 years and with an interest coverage ratio<1 over three consecutive years. Capital stock and employment refer to the share of capital and labour sunk in zombie firms. The sample excludes firms that are larger than 100 times the 99th percentile of the size distribution in terms of capital stock or number of employees. Source: OECD calculations based on ORBIS.

Figure: Zombie firms according to OECD definition.

Zombie firms

Figure: Zombie firms, alternative definitions.

Why should I care?

- Banks can be stuck in **zombie lending** (Peck and Rosengren, 2005; Caballero et al, 2008).
- **Crowding out** of financial resources, especially in times of crisis (Schivardi et al., 2020, 2022).
- **Lower aggregate productivity** by dragging down country averages (Mc Gowan et al., 2018);
- **Deter entry** of more productive firms, hence less competitive pressures on incumbents (see also discussion on reforms of bankruptcy laws).

Our contribution

- we propose **machine learning** techniques to predict **zombie firms**.
- we derive the risk of failure based on disclosed financial information and non-random missing values of 304,906 firms active in **Italy** from 2008 to 2017.
- we identify zombies as firms that persist in a state of financial distress, i.e., their forecasts fall into the risk category above a high-risk threshold for at least three consecutive years.
- we implement a gradient boosting algorithm (XGBoost) that exploits information about **missing values**.
- The inclusion of missing values in our predictive model is crucial because patterns of undisclosed accounts are correlated with firm failure.
- we show that our approach outperforms (i) proxy models such as Z-scores and the Distance-to-Default, (ii) traditional econometric methods, and (iii) other widely used machine learning techniques.

Literature review (1)

- Originally, 'zombie lending' (Caballero et al., 2008): **under capitalized banks** can decide to cut credit to more viable projects to avoid public disclosure of non-performing loans in their portfolio. The intuition is that 'zombie firms' receive **hidden subsidies** in the form of bank credit. See also Schivardi (2020) on crowding out of resources for healthy Italian firms in times of crisis.
- But what is a 'zombie'? The seminal working definition by Caballero (2008) is based on how present interest payments compared to an estimated benchmark of debt structure and market interest rate. Other proxy indicators by Bank of England (2013) are **negative value added** and profitability.
- McGowan et al. (2018) considers also misallocation of productive resources (not only financial): look at **productivity levels** and consider market entry/exit barriers (e.g. bankruptcy laws). See also a few discussion papers by OECD (2017a; 2017b).

Literature review (2)

Table 4 SL literature on firms' failure and financial distress

References	Domain	Output	Country, time	Data set size	Primary SL-method	Attributes	GOF
Alaka et al. [2]	CS	Bankruptcy	UK (2001–2015)	30,000	NN	5	88% (AUC)
Barboza et al. [9]	CS	Bankruptcy	USA (1985–2014)	10,000	SVM, RF, BO, BA	11	93% (AUC)
Bargagli-Stoffi et al. [12]	ECON	Fin. distress	ITA (2008–2017)	304,000	BART	46	97% (AUC)
Behr and Weinblat [14]	ECON	Bankruptcy	INT (2010–2011)	945,062	DT, RF	20	85% (AUC)
Bonello et al. [17]	ECON	Fin. distress	USA (1996–2016)	1848	NB, DT, NN	96	78% (ACC)
Brédart [18]	BMA	Bankruptcy	BEL (2002–2012)	3728	NN	3	81%(ACC)
Chandra et al. [23]	CS	Bankruptcy	USA (2000)	240	DT	24	75%(ACC)
Cleofas-Sánchez et al. [25]	CS	Fin. distress	INT (2007)	240–8200	SVM, NN, LR	12–30	78% (ACC)
Danenas and Garsva [30]	CS	Fin. distress	USA (1999–2007)	21,487	SVM, NN, LR	51	93% (ACC)
Fantazzini and Figini [36]	STAT	Fin. distress	DEU (1996–2004)	1003	SRF	16	93% (ACC)
Hansen et al. [71]	ECON	Fin. distress	DNK (2013–2016)	278,047	CNN, RNN	50	84% (AUC)
Heo and Yang [47]	CS	Bankruptcy	KOR (2008–2012)	30,000	ADA	12	94% (ACC)
Hosaka [48]	CS	Bankruptcy	JPN (2002–2016)	2703	CNN	14	18% (F-score)
Kim and Upneja [54]	CS	Bankruptcy	KOR (1988–2010)	10,000	ADA, DT	30	95% (ACC)
Lee et al. [63]	BMA	Bankruptcy	KOR (1979–1992)	166	NN	57	82% (ACC)
Liang et al. [65]	ECON	Bankruptcy	TWN (1999–2009)	480	SVM, KNN, DT, NB	190	82% (ACC)
Linn and Weagley [66]	ECON	Fin. distress	INT (1997–2015)	48,512	DRF	16	15% (R^2)
Moscatelli et al. [77]	ECON	Fin. distress	ITA (2011–2017)	250,000	RF	24	84%(AUC)
Shin et al. [88]	CS	Bankruptcy	KOR (1996–1999)	1160	SVM	52	77%(ACC)
Sun and Li [91]	CS	Bankruptcy	CHN	270	CBR, KNN	5	79% (ACC)
Sun et al. [92]	BMA	Fin. distress	CHN (2005–2012)	932	ADA, SVM	13	87%(ACC)
Tsai and Wu [94]	CS	Bankruptcy	INT	690–1000	NN	14–20	79–97%(ACC)
Tsai et al. [95]	CS	Bankruptcy	TWN	440	ANN, SVM, BO, BA	95	86% (ACC)
Wang et al. [99]	CS	Bankruptcy	POL (1997–2001)	240	DT, NN, NB, SVM	30	82% (ACC)
Udo [96]	CS	Bankruptcy	KOR (1996–2016)	300	NN	16	91% (ACC)
Zikeba et al. [105]	CS	Bankruptcy	POL (2000–2013)	10,700	BO	64	95% (AUC)

Abbreviations used—Domain: ECON: Economics, CS: Computer Science, BMA: Business, Management, Accounting, STAT: Statistics. Country: BEL: Belgium, ITA: Italy, DEU: Germany, INT: International, KOR: Korea, USA: United states of America, TWN: Taiwan, CHN: China, UK: United Kingdom, POL: Poland. Primary SL-method: ADA: AdaBoost, ANN: Artificial neural network, CNN: Convolutional neural network, NN: Neural network, GTB: gradient tree boosting, RF: Random forest, DRF: Decision random forest, SRF: Survival random forest, DT: Decision Tree, SVM: support vector machine, NB: Naive Bayes, BO: Boosting, BA: Bagging, KNN: k-nearest neighbor, BART: Bayesian additive regression tree, DT: decision tree, LR: Logistic regression. Rate: ACC: Accuracy, AUC: Area under the receiver operating curve. The year was not reported when it was not possible to recover this information from the papers

Figure: Bargagli-Stoffi, F. J., Niederreiter, J., & Riccaboni, M. (2021).

Machine Learning in Economics

Machine learning techniques have been applied, so far, to a variety of economic problems Mullainathan and Speiss (2017):

- Generation of **new data sets** (Jean et al., 2016; Cavallo and Rigobon, 2016).
- **Prediction** (Bajari et al., 2015; Kleinberg et al, 2015; Kleinberg et al., 2017).
- **Testing theory** (Hatford et. al., 2016; Erev et al., 2017; Plonsky et al., 2017).
- **Causal Inference** (Hill, 2011; Belloni et al., 2011, 2014; Athey and Imbens, 2016)

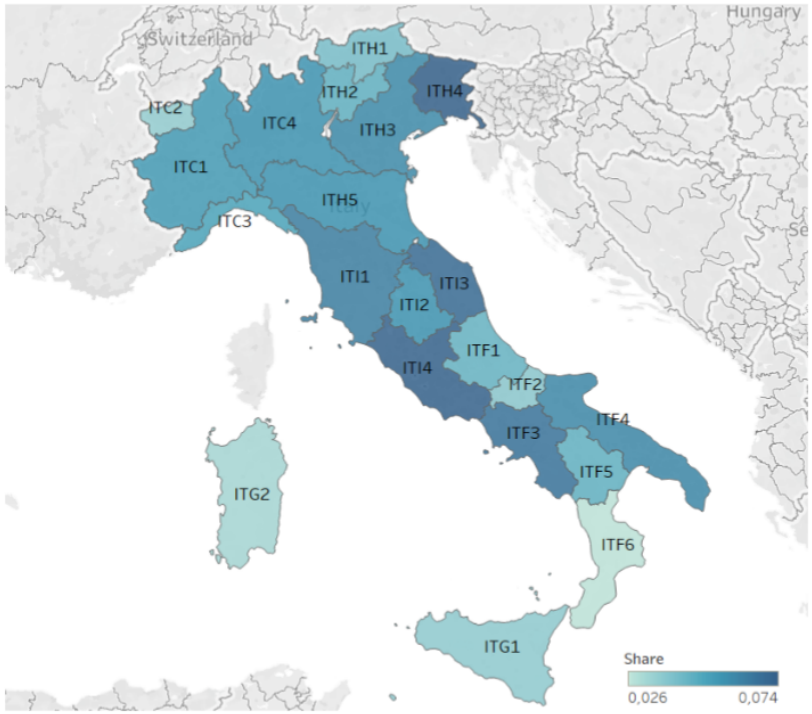
Machine Learning in Economics

Machine learning techniques have been applied, so far, to a variety of economic problems Mullainathan and Speiss (2017):

- Generation of **new data sets** (Jean et al., 2016; Cavallo and Rigobon, 2016).
- **Prediction** (Bajari et al., 2015; Kleinberg et al, 2015; Kleinberg et al., 2017).
- **Testing theory** (Hatford et. al., 2016; Erev et al., 2017; Plonsky et al., 2017).
- **Causal Inference** (Hill, 2011; Belloni et al., 2011, 2014; Athey and Imbens, 2016)

Data

- We train our algorithm on 304,869 manufacturing firms in Italy active in the period 2008-2017 with at least a value known for sales/turnover. The original source is Orbis, by Bureau Van Dijk.
- For each firm we have a status with a status precise date.



Status	Active	Bankrupted	Dissolved	In Liquidation	Total
Sample	287,586	1,533	8,540	7,221	304,906
Percentage	94.33%	0.50%	2.80%	2.37%	100%

Missing values (1)

Missing predictor	Odds ratio	Std. Error	N. obs.	Pseudo R^2
Interest Coverage Ratio	5.70***	(1.07)	298,873	0.051
Interest Benchmarking	4.09***	(0.75)	298,873	0.043
Negative Value Added	6.65***	(1.22)	298,873	0.052
Z-score	10.29***	(2.21)	298,873	0.069
Total Factor Productivity	7.04***	(1.22)	298,873	0.056
Profitability	5.70***	(1.07)	298,873	0.051

- Odds ratios according to a logit specification in which the dependent variable is a firm failure and the binary regressor equals one if at least one missing value was found in the last three years. Fixed effects at the region and industry level. Errors are clustered by industry.
- In our sample, we find that about 19% of firms have an ICR smaller than one, but at the same time, there are 62.50% firms whose ICR information is not available at all!

Missing values (1)

Table A.2: Missing predictors and firms' failures - Chi-square tests

	Firm's failure				Test Statistic
	0		1		
	$N = 287587$		$N = 17319$		
Interest Benchmarking : 0	38%	(110524)	61%	(10530)	χ^2_1=3414.25, P<0.001
Interest Benchmarking : 1	62%	(177063)	39%	(6789)	
Interest Coverage Ratio : 0	37%	(105907)	49%	(8422)	χ^2_1=970.93, P<0.001
Interest Coverage Ratio : 1	63%	(181680)	51%	(8897)	
Negative value added : 0	34%	(98014)	63%	(10915)	χ^2_1=5958.81, P<0.001
Negative value added : 1	66%	(189573)	37%	(6404)	

Note: Chi-square tests for the null hypothesis that missing predictors do not correlate with the event of failure. Number of observations in parentheses.

Missing values (2)

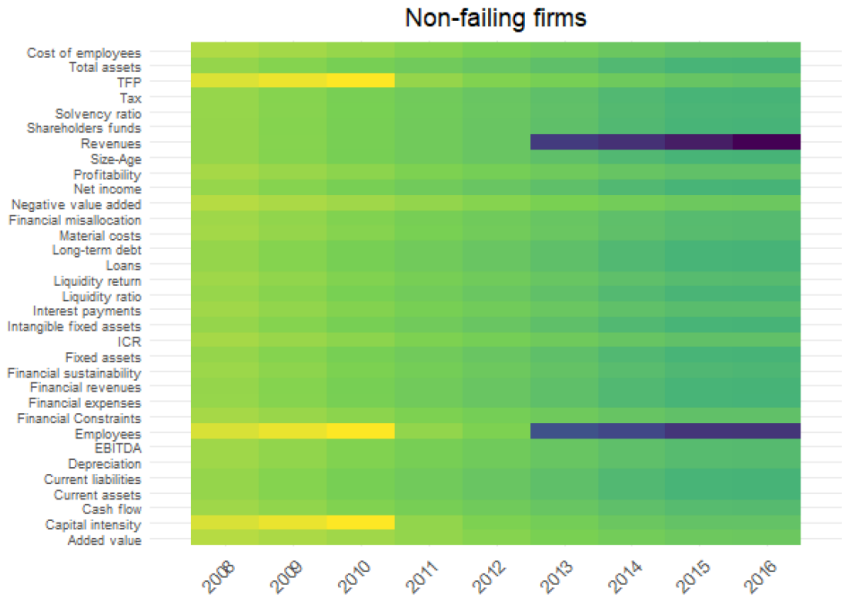


Figure: Panel (A): Share of missing values out of 287,787 non-failing firms

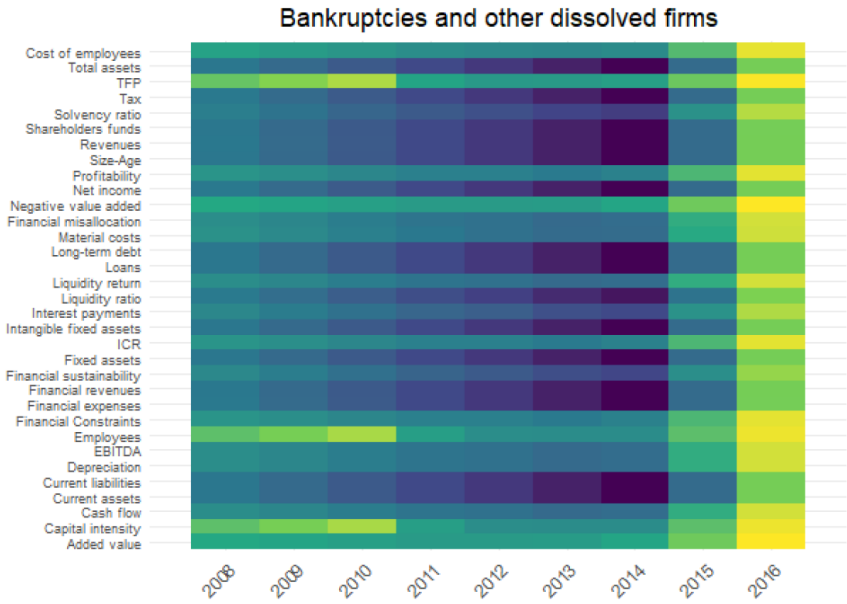
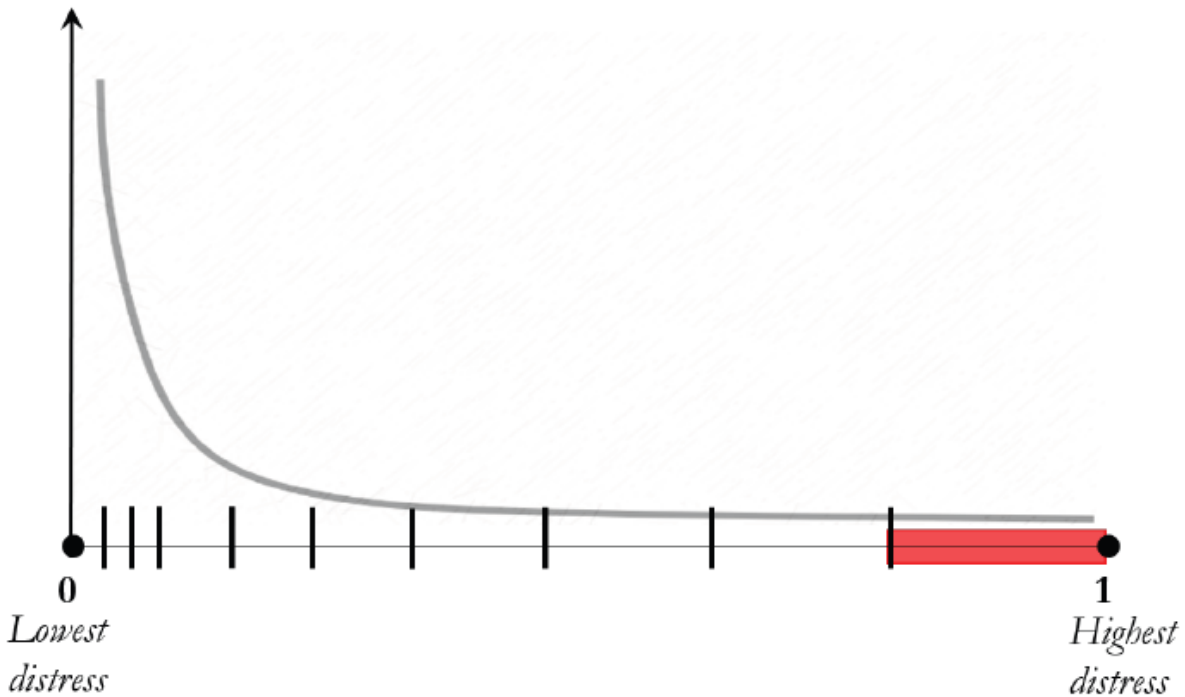


Figure: Panel (B): Share of missing values out of 4,718 bankruptcies and other dissolved firms.

Financial accounts show different patterns of missing values across firms (not at random, Chi-square tests): **firms may avoid disclosure relatively more when in trouble**

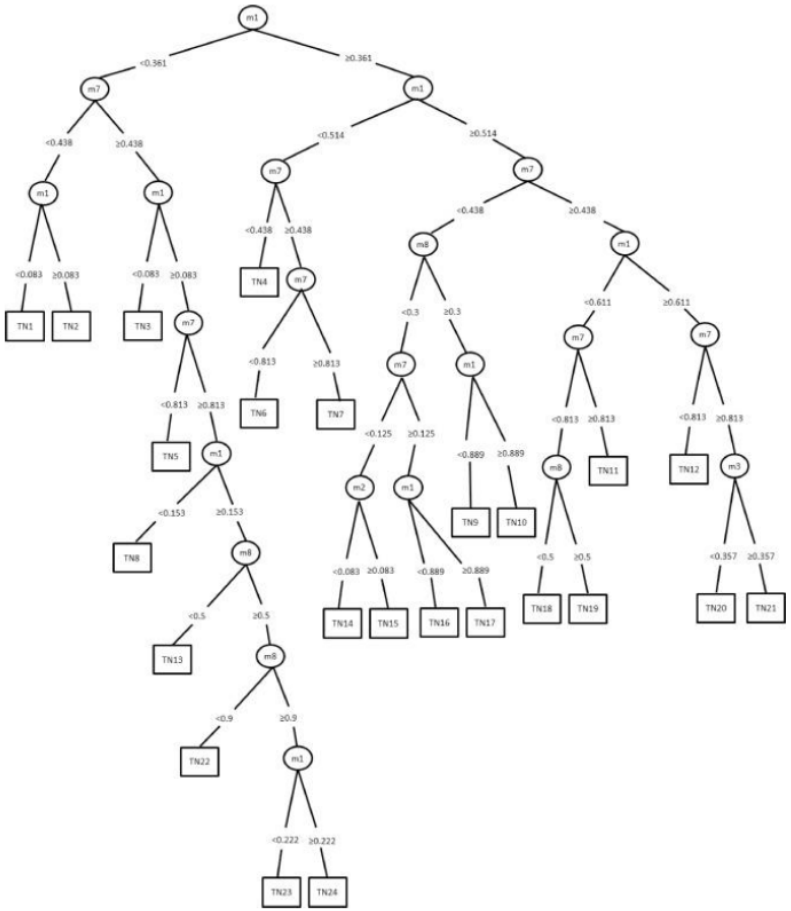
Empirical strategy

We use past information about already failed firms to assess what the probability is that another firm in a similar shape will go bankrupt (Kleinberg et al., 2015)



Bayesian Additive Regression Trees

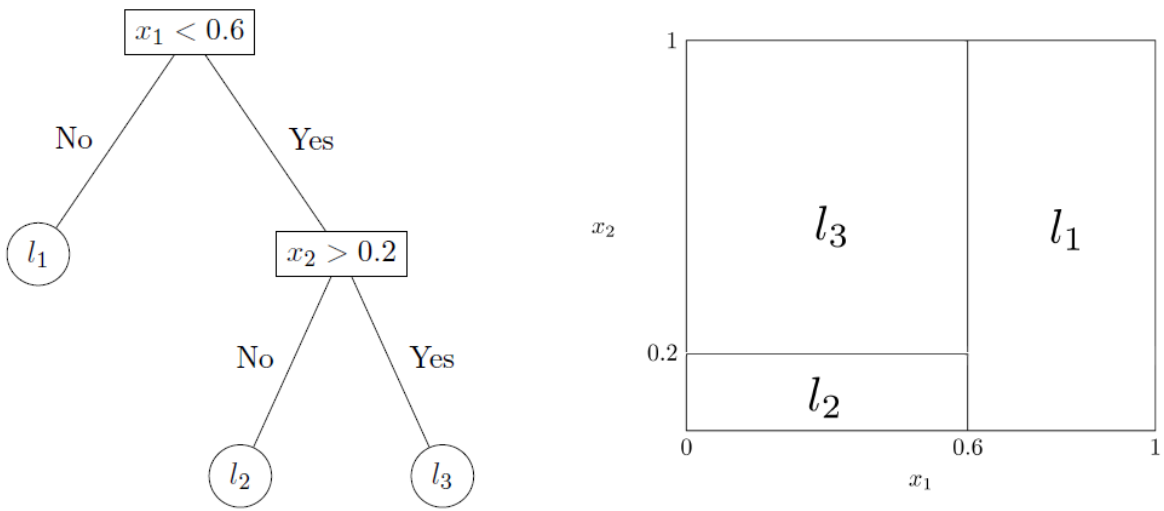
Bayesian additive regression trees (BART) provides a flexible approach to fitting a variety of regression models while **avoiding strong parametric assumptions**. The sum of trees model is embedded in a Bayesian inferential framework to support uncertainty quantification and provide a principled approach to regularization through prior specification (Hill et al., 2019)



Classification and regression trees

Decision tree-based algorithms are considered suitable tools in these cases due to their flexibility and high performance. The Classification and Regression Tree (CART) algorithm, first introduced by Breiman et al. (1984), is a widely used decision tree algorithm that constructs binary trees where each node is divided into only two branches.

Figure 3: An example of binary tree



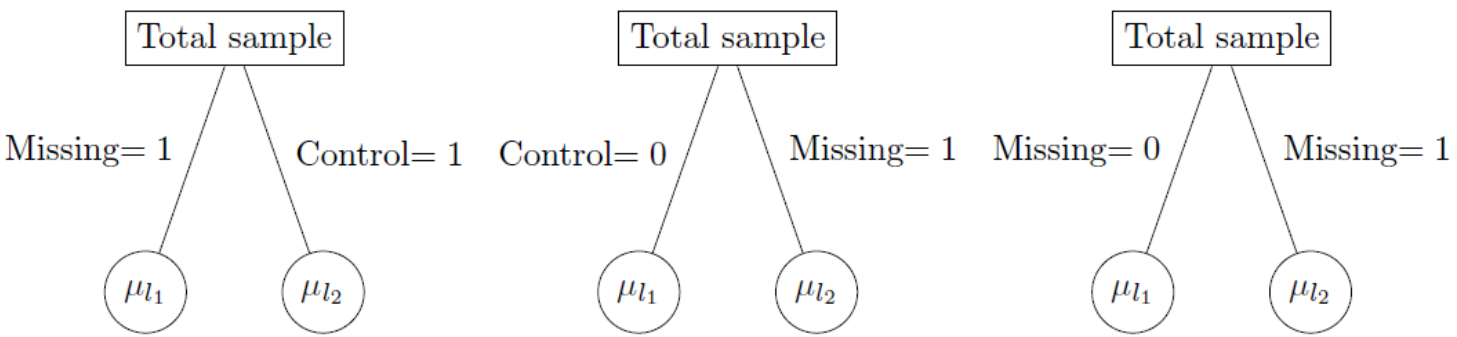
(a) Binary tree

(b) Feature space

In Figure (3a), the internal nodes are labeled by their splitting rules and the terminal nodes by the corresponding parameters l_i. Figure (3b) shows the corresponding partition of the feature space.

BART-MIA

- **BART-MIA** extends the original BART algorithm by incorporating additional information coming from patterns of missing values (Kapelner and Bleich, 2015).
- This is done by introducing, in each binary tree component of the BART algorithm, the possibility of **splitting on a missingness feature**.
- This splitting rule allows trees to better capture the direct influence of missing values as a further predictor of the response variable (Twala et al., 2008).



The three potential trees from the MIA procedure in a simple case with only one binary variable (Control $\in \{0, 1\}$).

XGBoost

XGBoost is a gradient-boosting algorithm. The algorithm uses a standard boosting method where J decision trees are sequentially created to approximate the outcome. Each tree uses the information learned from the previous trees, and the final model can be expressed as follows:

$$Y_{i,t} = \sum_{j=1}^{J} \mathcal{T}_j \left({}_{i,t-1}; \mathcal{D}_j, \mathcal{W}_j \right) + \epsilon_{i,t-1}$$

where $\mathcal{T}_j \left({}_{i,t-1}; \mathcal{D}_j, \mathcal{W}_j \right)$ corresponds to an independent tree with structure $\mathcal{D}_j$ and leaf weights $\mathcal{W}_j$. Note that $\epsilon_{i,t-1}$ is typically assumed to be zero-mean, but no probabilistic assumptions are made about it. The model approximation is built additively, minimizing the loss function iteratively. The loss function includes a regularization term to penalize the complexity of the model and avoid overfitting, and has the following form:

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N} L\left(\hat{Y}_{i,t}, Y_{i,t}\right) + \sum_{j=1}^{J} \Omega\left(\mathcal{T}_j\right) \\ \Omega(\mathcal{T}_j) &= \gamma T_j + \frac{1}{2}\lambda \|\mathcal{W}_j\|^2 \end{aligned}$$

where T_j and $\mathcal{W}_j$ represent the number and weights of the leaves of the j-th tree, respectively, while γ and λ are regularization parameters used to reduce complexity and avoid overfitting.

Results

Table 3: Models' horse race: performance measures

(a) Complete-Case analysis

Method	AUC	PR	F1-Score	BACC	R^2	Time
Logit	0.8966	0.4542	0.1833	0.7504	0.2658	9.13
Ctree	0.8957	0.4444	0.1987	0.7668	0.2640	572.46
Random Forest	0.9117	0.5233	0.1907	0.7595	0.3135	261.62
CC-XGBoost	0.9140	0.5170	0.1833	0.7504	0.3126	43.66
BART	0.9185	0.5221	0.1843	0.7533	0.3179	1249.05
Super Learner	0.9231	0.5464	0.1844	0.7535	0.3373	4147.87

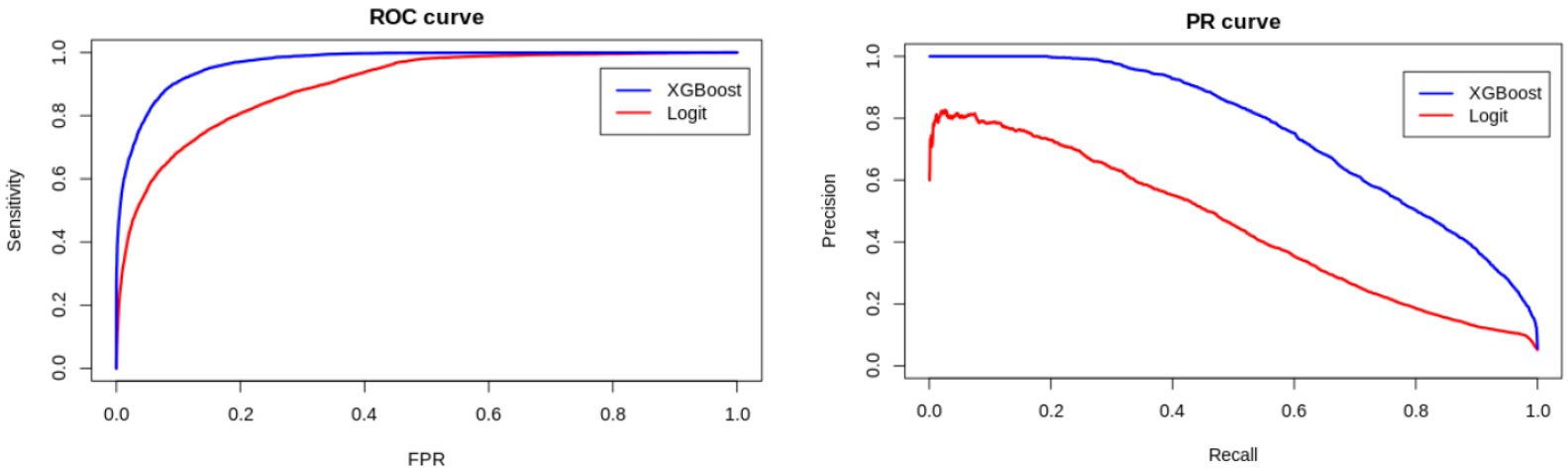
(b) Missing-Aware analysis

Method	AUC	PR	F1-Score	BACC	R^2	Time
XGBoost	0.9685	0.7591	0.2070	0.7646	0.5243	24.70
BART-MIA	0.9681	0.7516	0.2092	0.7676	0.5178	1126.88

All algorithms are trained with five-fold cross-validation. The training and test sets include 95,970 and 19,194 observations in each iteration, respectively. All metrics correspond to the five-fold average. Time indicates the average seconds required to train the model in each fold.

Goodness of fit

Figure 4: Goodness-of-Fit Scores



The ROC and PR curves for the Complete-Case Logit and the XGBoost models. Each plot shows the five-fold cross-validated mean curves, with the mean taken with respect to the ROC and PR curves of each validation set along the cross-validation routine.

Validation vs proxy models of credit scoring

Table 4: Goodness-of-fit: Distance-to-Default (DtD), Z-scores and XGboost

	DtD		Z-Scores		XGboost	
Percentile	Precision	FDR	Precision	FDR	Precision	FDR
1	0.3314	0.6686	0.2239	0.7761	0.9850	0.0150
2	0.3314	0.6686	0.2102	0.7898	0.9054	0.0946
3	0.3314	0.6686	0.2070	0.7930	0.8316	0.1684
4	0.3314	0.6686	0.1986	0.8014	0.7517	0.2483
5	0.3020	0.6980	0.1937	0.8063	0.6715	0.3285
6	0.2723	0.7277	0.1882	0.8118	0.6037	0.3963
7	0.2497	0.7503	0.1875	0.8125	0.5447	0.4553
8	0.2334	0.7666	0.1831	0.8169	0.5018	0.4982
9	0.2226	0.7774	0.1769	0.8231	0.4600	0.5400
10	0.2139	0.7861	0.1745	0.8255	0.4312	0.5688

Shapley values (1)

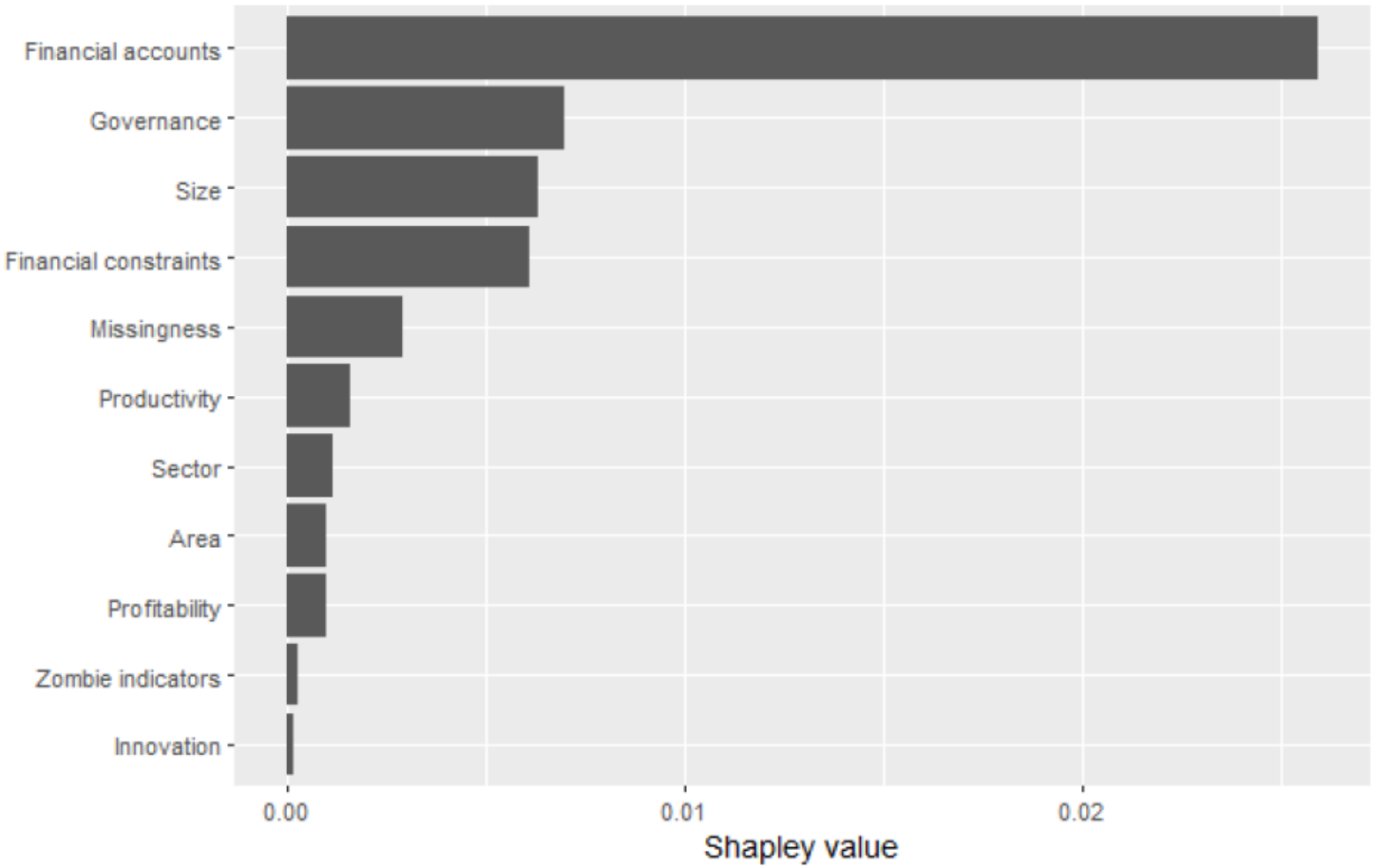
- Assume that S is a q-dimensional subset of variables, m is a generic variable in S ($m \subset S$), and $v(T)$ is a generic value function that takes in the subset S and returns real-valued payoff of the model (e.g., the goodness-of-fit) created using S or subsets thereof.
- Then the Shapley value $\phi_m(v)$ for a generic variable m is:

$$\phi_m(v) = \frac{1}{q} \sum_{S \subseteq \{1,\ldots,q\} \setminus \{m\}} [v(S \cup \{m\}) - v(S)] \frac{|S|!(q - |S| - 1)!}{q!}. \quad (1)$$

- Using (1), we see how the Shapley value is computed by calculating a weighted average gain in payoff (read: gain in goodness-of-fit) that the variable m yields when included in all subsets of variables that exclude m.

Shapley values (2)

Figure 6: Shapley values for the groups of variables in the predictive model



A case for zombie firms (1)

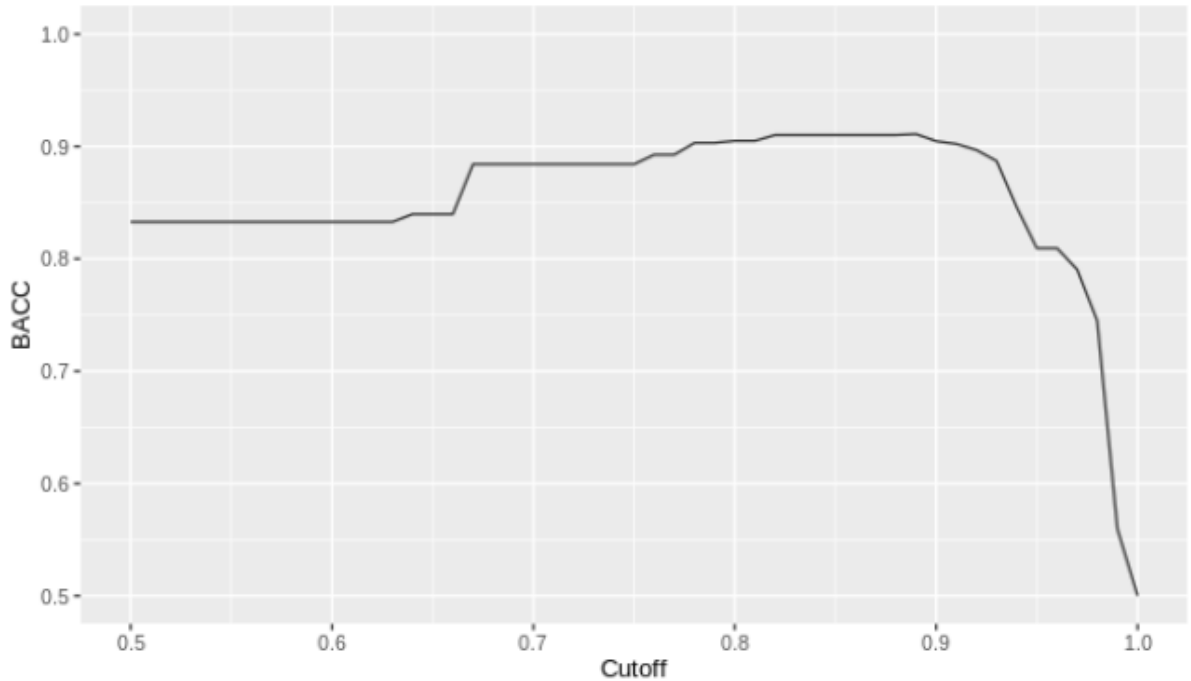
- a viable firm does not easily shift into financial distress, but if it does, it is difficult to recover from it.
- it makes sense to set an appropriate threshold that realistically reflects the most difficult situations in business life.
- we obtain this threshold by determining the cutoff that minimizes the combination of false positive and negative rates.
- the BACC, since it corresponds to a convex combination of the true positive and negative rates, which in turn are complementary to the false positive and negative rates.

Table 5: Transitions across deciles of risk

t / $t+1$	9th decile $t+1$	8th decile $t+1$	7th decile $t+1$	6th decile $t+1$	Below 6th decile $t+1$	Total $t+1$
9th decile t	0.46	0.22	0.12	0.08	0.12	1.00
8th decile t	0.22	0.22	0.17	0.13	0.26	1.00
7th decile t	0.11	0.16	0.19	0.15	0.39	1.00
6th decile t	0.07	0.11	0.15	0.18	0.49	1.00
Below 6th decile t	0.03	0.04	0.06	0.09	0.78	1.00

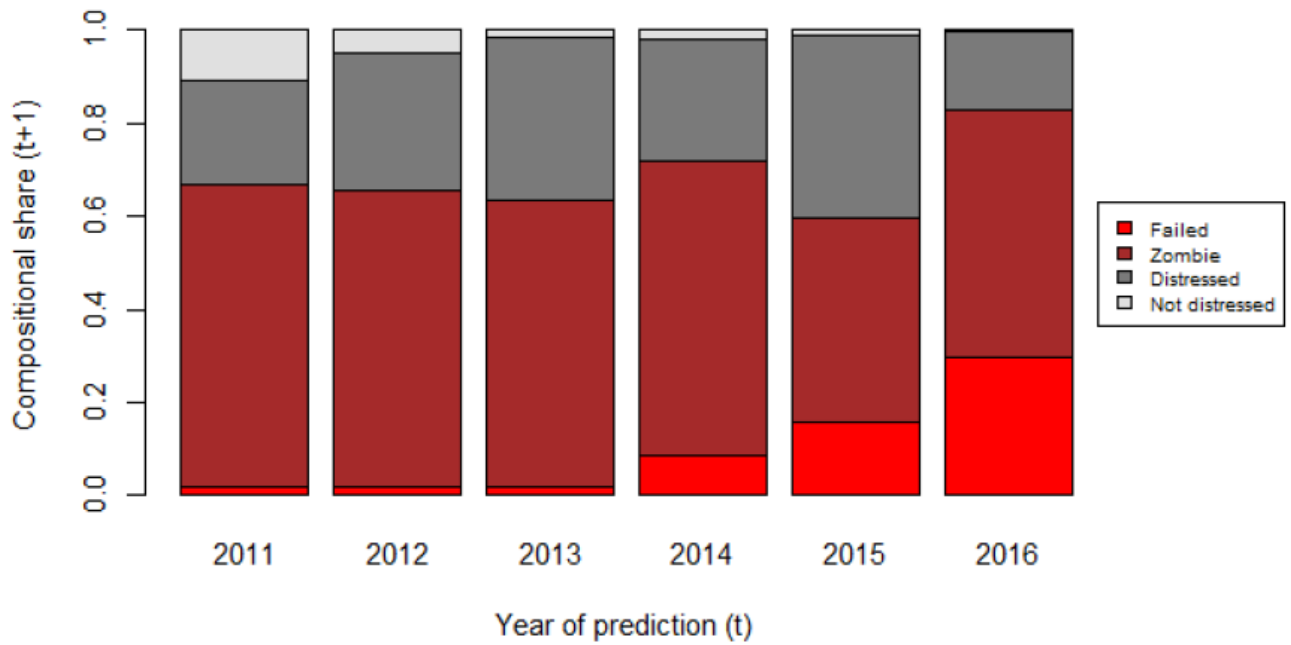
A case for zombie firms (2)

Figure 7: BACC at different cutoffs along the distribution of predictions



A case for zombie firms (3)

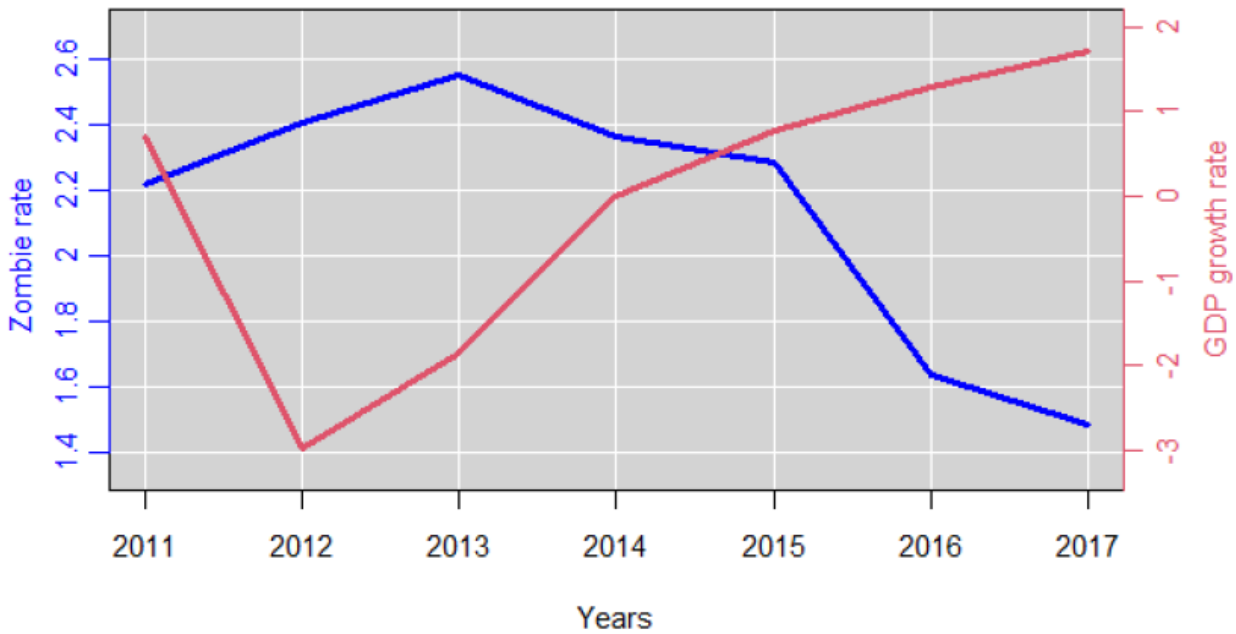
Figure 8: Transitions after predictions of a *zombie* status



The bars of the diagram show the transition of *zombie* firms in the years following the prediction: i) to failure (light red); ii) to remaining in a *zombie* status (dark red); iii) to relatively lower distress, i.e. between the 6th and 9th deciles (dark gray); iv) to a range of no distress, i.e., below the 6th decile (light grey). Note that we cannot report 2017 because we cannot compare it to actual observations in subsequent years.

Zombie firms in Italy

Figure 9: *Zombie firms* and the economic cycle



The share of *zombies* on the left axis is compared with nominal GDP growth rates on the right axis, obtained from the World Bank for the period 2011-2017. *Zombie firms* are firms that are at the right end (9th decile) of the predicted risk distribution for at least three consecutive years.

Productivity and size of zombie firms

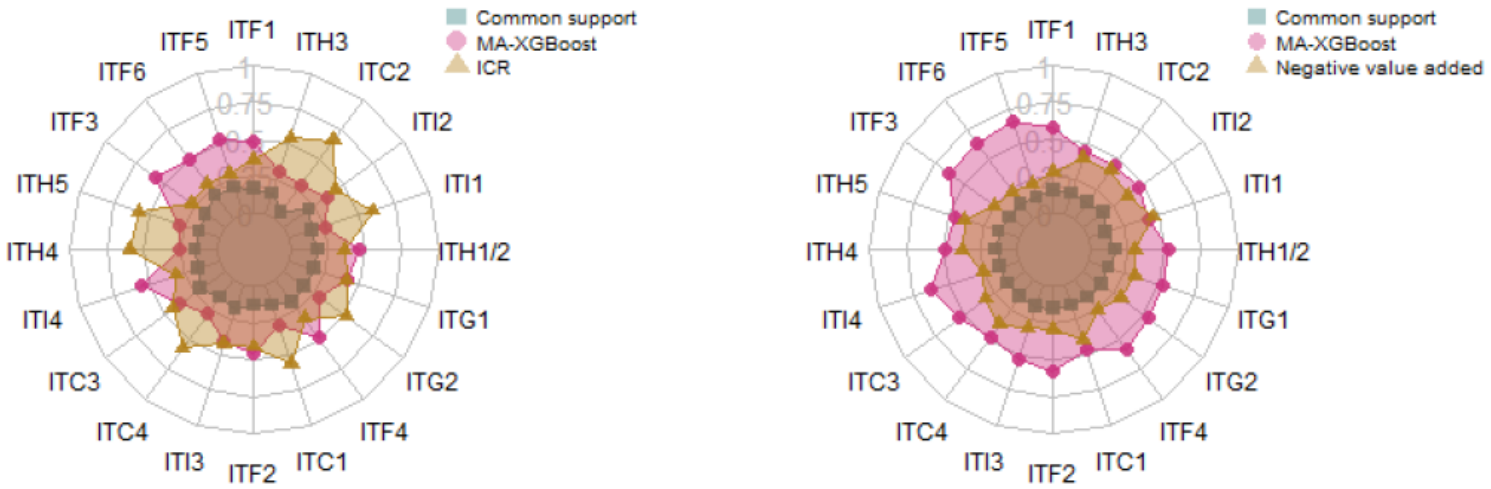
Table 6: Productivity and size premia for *zombies* vs. healthy firms

Indicator (in logs)	Coeff.	Std. Error	N. obs.	Adj. R squared
Total Factor Productivity	-0.197***	(.037)	600,771	.967
Labor Productivity	-0.450***	(.014)	559,315	.110
Sales	-2.066***	(.053)	1,234,750	.116
Employees	-0.170***	(.039)	1,119,486	.157

The table shows the coefficients of the linear models for panel data with fixed effects at the region and industry levels. We use pooled OLS estimation with cluster-robust standard errors to account for possible correlations within regions and industries. The dependent variable is a measure of firm productivity or size. The main covariate is an indicator that takes the value of one if the firm is classified as *zombie* in a given year of our sample, and zero otherwise. *Zombie firms* are defined as firms that are at the right end of the predicted risk distribution (above the 9th decile) for three consecutive years.

The regional distribution of zombie firms

Figure 10: *Zombie firms* and geography



(a) *Zombies* and interest payments

(b) *Zombies* and value destruction

Note: The rays of the radar show, at the regional level (NUTS 2-digit), the proportion of *zombie* firms versus firms that have an Interest Coverage Ratio (ICR) of less than one (panel a) and versus firms that have negative value added (panel b). The square nodes indicate the common areas where the segments overlap. The circles represent the fraction of *zombies* that we detect using XGboost. The triangles indicate the fraction of firms identified with $ICR < 1$ or negative value added. Along each ray, the values of the squares, circles, and triangles sum to one. Panel (a) and (b) include a total of 30,380 and 24,351 observations, respectively.

Conclusions

- **Machine learning** can derive non-trivial information on a battery of financial indicators and missing values, to successfully classify firms in risk categories after training on past failures.
- We classify as **zombie firms** the ones that persist in high-risk status because they are located on the right tail of our predictions for at least three years, beyond the 9th decile of risk, where we find that the chances to recover to smaller distress are minimal.
- In the **post-Covid scenario** the problem of separating the companies that can stay on their feet alone from the ones that conceal their insolvency is crucial.