IMT School for Advanced Studies, Lucca
Lucca, Italy

Machine Learning Firm Dynamics: Failure, Success, and Growth

PhD Program in Systems Science - Economics, Network, Business Analytics

XXXVI Cycle

By

Fabio Incerti

2025

The dissertation of Fabio Incerti is approved.

PhD Program Coordinator: Massimo Riccaboni, IMT School for Advanced Studies Lucca

Advisor: Massimo Riccaboni, IMT School for Advanced Studies Lucca

The dissertation of Fabio Incerti has been reviewed by:

Marco Grazzi, Department of Economic Policy, Università Cattolica del Sacro Cuore

Jerman Kaminski, Maastricht University, School of Business and Economics

IMT School for Advanced Studies Lucca
2025

Contents

Acknowledgements

These past four years have been a time of tremendous personal evolution and growth. I would like to express my gratitude to all those who contributed to completing this thesis, starting with my supervisor, Massimo Riccaboni.
A special thank you goes to Mark Kattenberg, for the kindness, assertiveness, and deep respect that have shaped our collaboration, as well as for his ability to approach even the most challenging situations with a positive spirit.
I am deeply grateful to my family and friends, who—whether near or far—have always made me feel a strong sense of home. Thanks also to all the colleagues I met along the way at the IMT School, with whom I shared this journey. Finally, a heartfelt thank you to Vittorio Mattei, for the great friendship born through this PhD experience.

Introduction

The thesis explores the application of cutting-edge, tree-based machine-learning techniques to analyze and predict the dynamics of firm failure, success, and growth. It is structured around three key areas: the detection of non-viable firms, the analysis of factors driving the future success of startups, and the examination of firm growth dynamics while addressing methodological limitations. Given the endemic missingness affecting firm-level data—often highly informative in prediction—particular attention is dedicated to the handling and usage of missing values within predictive frameworks. By integrating predictive analytics with empirical insights, the thesis seeks to enhance understanding and transparency in firm dynamics, offering insights for both academic research and practical applications.

Chapter 1 introduces a framework for identifying zombie firms, i.e., companies that persist in a high-risk status despite financial distress. Using a machine-learning approach, we define zombies based on predictive thresholds and analyze their characteristics, distribution, and economic impact. A key focus is the role of missing financial data, which we incorporate into the predictive model to enhance accuracy. Leveraging a rich dataset of Italian firms, we evaluate the effectiveness of various machine-learning methods in forecasting firm viability and compare their performance with traditional credit risk models. The findings provide insights for policymakers and stakeholders aiming to improve financial stability and resource allocation.

Chapter 2 investigates the determinants of startup success, focusing

on whether survival and long-term performance are driven by initial conditions or by firms' ability to adapt and develop capabilities over time. The analysis is framed within two theoretical perspectives: noisy learning and capability learning. The former suggests that firms discover if they have viable business models through experience, while the latter posits an adaptive learning process where firms actively improve their business model and strategies over time. Using a comprehensive dataset of newly founded firms in the Netherlands, the chapter employs gradient boosting to assess the extent to which early-life firm characteristics predict future success and interpret the results in term of noisy and capability learning theories. The findings provide insights into startup dynamics and offer policy recommendations aimed at fostering the emergence of high-growth and high-productivity firms.

Chapter 3 examines firm growth dynamics, a key topic in economic research, particularly regarding the factors that drive growth and the validity of Gibrat's law, which posits that firm growth is independent of size. However, empirical studies face significant challenges due to sample attrition and missing data, which can bias estimates and obscure the true relationship between firm growth, size, and innovation. To address these issues, this chapter proposes an extension of Heckman's two-step procedure, integrating machine-learning techniques to improve the correction for selection bias in firm growth analysis. By replacing the traditional probit model with a more flexible predictive algorithm, this approach enhances the ability to account for complex selection mechanisms and missing data patterns. The study focuses on the role of innovation—measured through intellectual property indicators such as patents and trademarks—as a driver of growth. The findings contribute to the debate on Gibrat's law and underscore the advantages of combining econometric methods with machine learning to obtain more accurate and reliable estimates of firm growth dynamics.

Chapter 1

Machine-learning for zombie hunting: predicting distress from firms' accounts and missing values

This chapter is co-authored with Falco J. Bargagli-Stoffi, Massimo Riccaboni, and Armando Rungi, and was published in 2023 in the journal Industrial and Corporate Change.

1.1 The conceptual framework

The first chapter proposes machine-learning techniques as suitable tools to make predictions about business failure when information about firm viability is partially undisclosed. Therefore, we define *zombies* as firms that persist in a high-risk status because their predicted probability of failure is above a threshold for which a combination of the false positive rate (i.e., predicting a firm that did not fail as failed) and the false negative rate (i.e., predicting a firm that failed as not failed) is minimized. We

prove that the latter is also the distributional segment in which the probability of transitioning to a lower risk of failure is minimal. According to our machine-learning approach, the Italian proportion of *zombie* firms in the analysis period is between 1.5% and 2.5%. *Zombie* firms are, on average, 19% less productive and 17% smaller than viable firms. Interestingly, we find that *Zombies* are countercyclical, as their share increases in times of crisis and decreases in economic recovery.

The problem of identifying non-viable firms is important to scholars and practitioners, whether to assess the credit risk of an individual firm or to identify the part of an entire economy that is in trouble. Originally, the notion of *zombie* firms was associated with the phenomenon of "*zombie lending*", in which banks extend credit to otherwise insolvent borrowers. In some cases, zombie lending is a deliberate strategy to avoid a bank's budget restructuring while only apparently complying with capital standards set by financial regulators (Bonfim et al., 2020), e.g., the case of Japanese banks in the 1990s (Peek and Rosengren, 2005; Caballero, Hoshi, and Kashyap, 2008). More recently, Schivardi, Sette, and Tabellini (2021) studied Italian companies during the 2008 financial crisis and found that the misallocation of credit under *zombie lending* increased the default rate of otherwise healthy firms while decreasing the default rate of non-viable firms. This was because undercapitalized banks could restrict lending to more viable projects to avoid disclosing non-performing loans in their portfolios. Paradoxically, severely distressed companies can appear resilient in times of financial crisis thanks to continued access to financial resources.

From a more general perspective, recent studies introduce *zombies* as firms that have persistent problems meeting their interest payments (Andrews, Muge Adalet McGowan, Millot, et al., 2017; Müge Adalet McGowan, Andrews, and Millot, 2018; Banerjee and Hofmann, 2018; Andrews and Petroulakis, 2019), thus expanding the original category to include firms that are in some form of financial distress. Based on proxy indicators from financial accounting, they show that *zombies* account for a non-negligible share of modern economies—up to 10% of incumbent firms—while absorbing up to 15%, 19%, and 28% of the capital stock in

countries such as Spain, Italy, and Greece, respectively. When market exit or restructuring delays occur, they drag down aggregate productivity by hindering the reallocation of resources in favour of healthier firms and preventing the entry of potentially more innovative and younger firms. It is therefore argued that identifying non-viable, financially distressed firms can be particularly useful in avoiding the misallocation of productive and financial resources.

Against this background, we argue that the empirical problem of assessing whether a firm is a *zombie* is closely related to the more general problem of determining its credit risk. Ultimately, a *zombie* is a non-viable firm that may escape bankruptcy despite its extreme financial distress—i.e., despite scoring the highest credit risk. From another perspective, healthier firms are the furthest from bankruptcy and *zombie* status. Traditionally, credit risk has been studied from the perspective of a financial firm, which must assess the health of a company using information, albeit limited, from financial books and public records.[1]. Thus, for decades, academics and practitioners have attempted to determine a firm's profitability after benchmarking exercises on firm-level indicators of financial constraints—e.g., in estimating *Z-Scores* (Altman, 1968; Altman et al., 2000), *Distance-to-Default* (*DtD*; Merton, 1974), or investment-to-cash-flow sensitivity ((Fazzari, R. G. Hubbard, and Petersen, 1988). However, information on corporate viability may be incomplete due to strategic disclosure of relevant information and simplified financial records for certain categories of unlisted companies.

Thus, we propose a machine-learning approach to predict credit risk and zombie status from incomplete financial accounts. To show the potential of our approach, we worked with a sample of 304,906 Italian firms over the period 2008–2017. Italy is a compelling example of a country that hosts a relevant share of inefficient firms that hinder the growth po-

[1]The seminal reference is to a departure from the Modigliani and M. H. Miller (1958) theorem, according to which capital structure should not be relevant to a company's value if there are no market frictions, including bankruptcy costs. Thus, a firm's ability to raise external financing should depend solely on the profitability of its investment projects. However, since financial market frictions cannot be eliminated, a firm's capital structure actually provides information about the profitability of the firm and its assets. See also the discussion of Rajan and Zingales (1995) for an international perspective

tential of the economy (Calligaris, Del Gatto, Hassan, G. I. P. Ottaviano, et al., 2018).

The underlying intuition is simple: based on the experience of firms that failed in previous periods, we derive predictions about the risk of failure of active firms. Each time we compare the observed outcomes with the predicted outcomes, the algorithm updates and reduces the prediction errors in the next periods after processing new "in-sample" information about the financial accounts. In the end, we obtain a probabilistic measure of the likelihood that a business will fail. Our machine-learning framework improves upon existing benchmark models by leveraging a rich set of firm-level economic and financial indicators that potentially contain diverse information about the firm's core economic activity and ability to meet financial obligations.

Importantly, we find that emerging patterns of missing financial accounts correlate with firm failure, possibly because managers are more likely to conceal accounts when they are in financial distress. We provide evidence that most missing variables are often those that have been used as proxies for *zombies* or financial constraints in the previous literature. Therefore, we implement our missing-aware methodologies incorporating patterns of undisclosed accounts and test whether a substantial improvement in prediction occurs when missing data are properly handled. Using the missing values information as another predictor of outcome, we show that the best predictive algorithm in our setting, eXtreme Gradient Boosting (XGBoost) (T. Chen and Guestrin, 2016), explains up to 0.97 of the area under the curve (AUC), and its precision-recall (PR) performance reaches 0.76. XGBoost outperforms credit scoring models (i.e., Z-score and DtD models), standard econometric methods (i.e., logistic regression), and other machine-learning techniques (i.e., classification and regression tree [CART] and random forests).

We argue that asymmetric and undisclosed (missing) information is ubiquitous in corporate financial accounts. Therefore, simply omitting records with missing values would severely impair the search for predictors of zombies and lead to a loss of precision and bias (Little and Rubin, 2019). If the missing information is correlated with firm failures, ex-

cluding these firms would significantly reduce the number of failures in the sample and thus hinder the algorithm's learning for these instances. This problem could potentially be avoided by a missing data imputation approach that would allow the inclusion of these firms in the training sample. On the other hand, the missing information per se constitutes relevant information to learn from failures. This may be the case if nonviable firms have the option of not disclosing information in their financial accounts. As a result, the mechanism in the data is missing not at random (MNAR), and the complete records are not a random sample of the population of interest. In this second case, if some companies consistently avoid disclosing part of their information, imputation of missing data would not be sufficient to solve the problem. Indeed, we know from previous literature that techniques for imputing data, such as mean or median imputation, can fail in the presence of MNAR because they distort the empirical distributions.[2]

Based on the previous considerations, we choose an approach that can simultaneously incorporate missing data into our analysis while being robust to MNAR. We implement two methods specifically designed to deal with nonrandom patterns of missing data: XGBoost and Bayesian Additive Regression Trees with Missing Incorporated in Attributes (BART-MIA) (Kapelner and Bleich, 2015). Although these methods differ in implementation, they have very similar routines for dealing with MNAR patterns. XGBoost uses default directions, also known as block propagation (Josse et al., 2019), to group all incomplete observations and send them to one side of the tree. The MIA method method (Twala, Jones, and Hand, 2008), extended by Kapelner and Bleich, 2015; Kapelner and Bleich, 2016, strengthens block propagation so that missingness can be used as an explicit feature to compute the best splits. According to Josse et al., 2019, MIA can handle both informative and noninformative missing values. Our results suggest that both XGBoost and BART-MIA effectively capture the direct influence of missing values as either implicit (block propagation) or explicit (MIA) predictors of the response variable.

[2]Further limitations of traditional approaches to imputing missing data are discussed in (He et al., 2010; T. K. White, Reiter, and Petrin, 2018; Little and Rubin, 2019).

Finally, the following analyses show that XGBoost has a significantly lower computational cost and relatively higher predictive power than BART-MIA. To shed more light on the information that contributes most to prediction, we use an interpretable framework that has its roots in game-theoretic Shapley values, introduced by Strumbelj and Kononenko (2010). Shapley values have recently been proposed to identify the economically meaningful nonlinearities learned by machine-learning models (Buckmann, Joseph, and Robertson, 2021). We find that no single financial indicator predicts failure better than the ensemble of predictors used in our machine-learning approach.

Economically informative groups of variables have heterogeneous predictive power. In particular, indicators of firms' financial constraint or previously used indicators of *zombies* are important. Nevertheless, they are of secondary importance when compared to information on firms' financial accounts, indicators of corporate governance, and the presence of nonrandom missing values. Our results confirm that machine-learning techniques perform better than single indicators when incorporating as much valuable in-sample information as possible while updating each time there is new out-of-sample information because prediction errors dynamically decrease after independent tests that minimize the discrepancies between the realized and predicted outcomes (Susan Athey, 2018).

Our framework is of particular interest to policymakers in designing optimal bankruptcy laws.[3] Tracking a company's bankruptcy risk allows all stakeholders, not just creditors, to understand whether there is an opportunity for restructuring and, if not, to prevent incumbent, albeit nonviable, firms from wasting additional economic resources. Evidence-based, interpretable methods are even more important after the recent pandemic crisis, as we believe that financial support must be targeted to companies that have a real chance of recovering and staying on their feet in normal times to avoid misallocation of resources.[4]

[3]See also the suggestions by the European Directive 2012/30/EU, and the recent Italian law on business failures on October 19, 2017, n. 155, which provides the legal basis for early notification of corporate crises to improve targeted interventions.

[4]For a first analysis of the impact of the pandemic crisis on Italian companies, see Schivardi and G. Romano (2020)

1.2 Data and preliminary evidence

We obtain the financial accounts from the ORBIS database[5], compiled by the Bureau Van Dijk, for manufacturing firms active in Italy for at least one year from 2008-2017. Italy is a compelling case to study business failure and *zombie firms*: it is a country where relatively inefficient firms hinder the economy's growth potential (Calligaris, Del Gatto, Hassan, G. I. Ottaviano, et al., 2016; Bugamelli et al., 2018), perpetuating geographic divergence (Rungi and Biancalani, 2019), and are studied extensively by international organizations (Müge Adalet McGowan, Andrews, and Millot, 2018; Andrews and Petroulakis, 2019). For our purpose, we use two main variables that help us identify business failure: the status of a firm and the date on which it becomes inactive. Table 1 shows our sample coverage by firm status over in the period of analysis. We assume that a firm has failed in the first year if it is reported as "bankrupt", "dissolved", or "in liquidation", as in the original data. Overall, the share of exiting firms accounts for about 5.7% of the total sample, which is close to the average official 6.3% obtained by ISTAT, the national statistics office, for the same period.

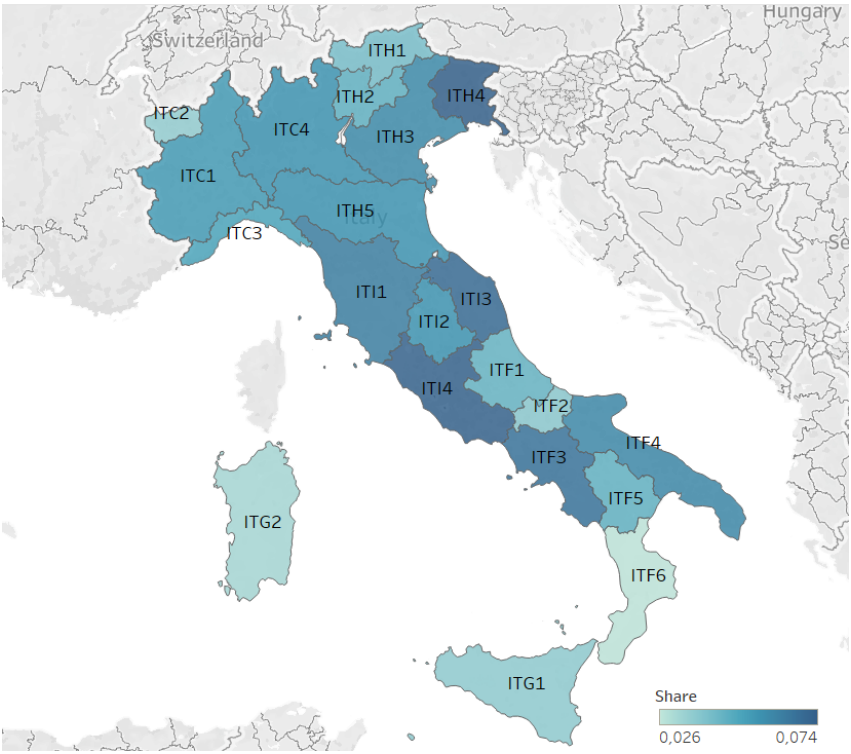
Figure 1 shows the proportions of firm failures by NUTS 2-digit regions. As expected, we find a higher concentration of failures in the north and center of the country, where economic density is higher. It is noteworthy that we fully represent the whole Italian territory since we detect business failures in every region during our period of analysis.

[5]ORBIS firm-level data (Orbis, 2020) have become a common source of global financial accounts. For previous use of this database, see Gopinath et al., 2017 and Cravino and Andrei A. Levchenko, 2016, among others. Coverage of smaller firms and some financial accounts may change across countries as national business registries impose different filing requirements, as observed in the validation exercises of Kalemli-Ozcan et al., 2015 and Gal, 2013. In the case of Italy, the original information provider for Italian financial accounts is CERVED, a credit rating agency. Bureau Van Dijk standardizes and translates the original financial accounts to make them comparable between countries. Note that, unlike other platforms of the same Bureau Van Dijk (e.g., AIDA or AMADEUS), ORBIS does not drop exiting firms in our analysis period. It supplements the financial accounts with other information from various sources on ownership, management and intellectual property rights, which we also use for predictions.

Table 1: Firms by status

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%

Figure 1: Geographic coverages



Notes: The proportion of business failures to the total number of firms for NUTS 2-digit regions. We assume that a firm fails when it is reported as bankrupt, dissolved, or in liquidation.

We use a set of economic and financial indicators to train our predic-

tive models. The battery of predictors includes (i) original financial accounts at the firm level; (ii) widely used indicators to proxy firm-level financial constraints; (iii) indicators previously used to detect zombie firms; (iv) indicators included as warnings of corporate crises in the recent Italian bankruptcy law.[6] Each predictor is detailed in Appendix Table A.1.1.

Note that many indicators we select as predictors have been used in various frameworks to assess the extent to which a firm is in trouble. From our machine-learning perspective, they cannot be interpreted as drivers of failure. It is sufficient that they contribute, albeit in a small way, to the assessment of a company's health. In our predictive framework, they might even border on multicollinearity, e.g., in the case of different measures of efficiency, liquidity, and solvency ratios. Since we are not interested in identifying a causal contribution to firm failure, high collinearity does not pose a problem for our predictions (Makridakis, Wheelwright, and Hyndman, 2008; Shmueli et al., 2010). On the contrary, we will discuss in Section 1.3.4 how, by construction, one cannot separate the empirical contribution of an indicator from the entire set in the context of a pure prediction problem. For this reason, the predictors we use should be considered as an inseparable ensemble, and a discussion of the statistical significance of the individual predictors is not relevant in our framework. In addition to mandatory and basic information (volume of activity, profits, location, industry affiliation, ownership, and intellectual property rights), many other financial accounts have different patterns of missing values over time. In the Appendix, Figure A.2.1 shows a map of missing values in our sample. The frequency of missing values affects all firms in the data, both active and failed and it is more concentrated in the financial accounts of recent years. In addition, liquidations exhibit less pronounced patterns over time than the other

[6] A recent reform of the bankruptcy law (L. 155/2017 and DL. 14/2019) proposes an early warning system based on indicators identified by practitioners, the purpose of which is to identify companies in distress in time to intervene to preserve entrepreneurial capabilities and find a way out of the crisis. It delegated practitioners (in particular the National Association of Chartered Accountants: Consiglio Nazionale dei Dottori Commercialisti e degli Esperti Contabili) to draw up a list of indicators that could help assess the state of crisis of a company

Table 2: Missing predictors and firms' failures

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 where the dependent variable is firm failure and the binary regressor equals 1 if at least one missing value is found in the last 3 years. Fixed effects at the region and industry levels are included. Errors are clustered by industry. Statistical significance at the 1% level is indicated by ***.

failure categories.[7] After running a series of χ^2 tests (see Appendix Table A.1.2), we find a positive statistical relationship between the patterns we observe in the sample and the event of a firm's failure. In short, a firm is more likely to fail if a pattern of missing financial accounts is observed. The exercise performed in Table 2 clearly shows such correlations. We run a simple logistic regression using the observed failure of a firm as the dependent variable. We then add a binary regressor equal to 1 if the predictor is missing at least once in the 3 years. Fixed effects by the NUTS 3-digit region and the NACE 2-digit industry are included. We report the results for each predictor per row in Table 2.

The above correlations are particularly relevant to the scope of our analyses. The main problem is sample selection when observations are selectively missing for some categories of firms. In this case, there are two potential sources of sample selection bias: (i) distressed firms vis-à-vis firms that are not in distress, as the former may have the incentive to disclose less information than the latter; (ii) smaller firms vis-à-vis bigger firms because the first are often exempted from a complete financial report, in accordance with Italian Regulation.[8] Of course, the two sets of

[7] There is no missing value for categorical variables (industry and region).

[8] Under Italian civil law, companies that do not list financial activities on the stock exchange have the option to provide more aggregate financial reports if their size does not

firms may overlap, as smaller companies may also be the ones that are proportionally more affected by financial difficulties. Using a missing-aware procedure, as described in Section 1.3, allows us to consider both sources of sample selection when patterns of missing financial accounts emerge since the algorithm considers such patterns as another predictor of firm failure.

Interestingly, we note that most of the missing variables are also those used in previous work as proxies for *zombies* or financial constraints. Take, for example, the case of the interest coverage ratio (ICR), which is derived as the ratio between a firm's earnings before interest and taxes (EBIT) and its interest expense. If the ICR is less than 1, Bank of England (2013) assumes that a company is a *zombie* because it is having trouble meeting its financial obligations. In our sample, we find that about 19% of firms have an ICR smaller than 1, but at the same time, there are 62.50% of firms whose ICR information is not available at all. Moreover a negative value added is an appropriate indicator for evaluating a *zombie* status, as it indicates that intermediate inputs have a higher market value than the firm's output. In the case of Italy, about 64.27% of enterprises do not report their value added in at least one period, while about 3% of them report a negative value. A negative value added is, of course, a more severe condition than a negative profit since a company can make no profits without destroying economic value. Indeed, firms' profitability is at the heart of two similar proxies for *zombies* used by Schivardi, Sette, and Tabellini (2021) when comparing firm-level profits to an external benchmark. Repeating the same exercises, we find that about 3% of Italian firms are distressed, while a large part of the sample (62.50%) does not report any information on the predictor. Finally, both Caballero, Hoshi, and Kashyap (2008) and Müge Adalet McGowan, Andrews, and Millot (2018) perform another benchmarking exercise, comparing the interest a firm pays to raise external finance with the cost opportunity to invest in alternative, safer investments. In our case, when we try to re-

simultaneously exceed two of the following thresholds in one or two consecutive periods: i) 4,400,000 euros in total assets; ii) 8,800,000 euros in operating revenues; iii) 50 employees. A simplified financial statement always includes the most significant items in the first or second position digit of the aggregation.

produce the same exercise with the yields of Italian government bonds with a ten-year maturity, we find that there is a high proportion (60.29%) of companies for which we have no information in at least one period in which they were active.

We also include an estimate of the total factor productivity (TFP) as a predictor, following the methodology proposed by Ackerberg, Caves, and Frazer (2015), to account for the simultaneity bias arising from ex post adjustments in combining factors of production. In this respect, firm-level TFP allows us to make predictions based on the ability to transform inputs and sell output in the market. Indeed, the relationship between financial constraints and productivity is one of the most debated issues (see, among others, Aghion, Bergeaud, et al., 2019; Ferrando and Ruggieri, 2018). The simple assumption for our forecasting models is that less productive firms are the ones that have more difficulty surviving in the market. At the same time, *zombie firms* have also often been defined in terms of (lack of) productivity (Müge Adalet McGowan, Andrews, and Millot, 2018; Andrews, Muge Adalet McGowan, Millot, et al., 2017; Andrews and Petroulakis, 2019; Schivardi, Sette, and Tabellini, 2020).[9]

1.3 Empirical strategy

It is difficult to identify non-viable companies for obvious reasons. If financial accounts are in bad shape, one could argue that it is only a matter of time before they become more competitive if conditions are right. If the balance sheets are good, one could argue that the worst is yet to come because bad management decisions will show up later. Trivially, only the already bankrupt companies were certainly not viable at some point. Still, an outside observer will never know when that happened because the manager of a company in trouble has an incentive not to disclose private information.

[9] Please note how T. K. White, Reiter, and Petrin (2018) recently highlighted the limitations of using missing imputation techniques for the variables used to calculate TFP in the US.

In principle, an analyst would like to observe the entire event horizon to discount all possible scenarios and understand the value of a company and its investment projects. However, this is not possible because it is the typical double problem of a financial institution facing uncertainty in the presence of information asymmetries. On the one hand, the company has a clear information advantage in its investment plans. On the other hand, both the financial institution and the company have a limited ability to predict future economic shocks, which can have either positive or negative effects.

In their seminal works, Altman (1968) and Ohlson (1980) apply standard econometric techniques—i.e., multiple discriminant analysis and logistic regression—to assess the probability of firm bankruptcy. Following these contributions and the Basel Accord II in 2004, default forecasts are based on standard reduced-form linear regression approaches. However, these approaches may fail because their limited complexity precludes nonlinear interactions among predictors, while their ability to handle large sets of predictors is limited due to potential multicollinearity problems.

Machine-learning algorithms compensate for these shortcomings by providing flexible models that allow nonlinear interactions in the space of predictors and the inclusion of a large number of predictors without the need to invert the covariance matrix of the predictors, thereby circumventing multicollinearity (Linn and Weagley, 2019). In addition, machine-learning models are directly optimized to perform the prediction task, resulting in better prediction performance in many complex situations.

The data science literature has already developed exercises to predict corporate failures using financial accounts but without a clear economic and financial framework. In this context, Falco J Bargagli-Stoffi, Niedermeiter, and Riccaboni (2021), Behr and Weinblat (2017), Linn and Weagley (2019), Moscatelli et al. (2019), De Martiis, Heil, and Peter (2020), Davies, Kattenberg, and Vogt (2023), Udo (1993), K. C. Lee, Han, and Kwon (1996), Tsai and Wu (2008), Sun and H. Li (2011), Brédart (2014), Tsai, Hsu, and Yen (2014), G. Wang, Ma, and S. Yang (2014), Alaka et al.

(2018), and Hosaka (2019).

With this in mind, we propose a machine-learning procedure that uses past information about previously failed firms to estimate the probability that another firm in a similar condition will go bankrupt. The wider the variety of past experiences on which we can draw, the more accurate the prediction about the health—or lack thereof—of a company (Kleinberg et al., 2015). Ultimately, our perspective is on a firm's (lack of) resilience, using potentially any observable data that might hold information about the firm's viability. In the end, we obtain a probabilistic measure at the firm level, ranging from 0 to 1, which tells us the probability that a firm will exit the market in the next period, given that other firms in a similar situation have done so. As shown in Figure 2, we can assess the distance of each company from the highest financial distress.

Let us consider a generic predictive model in the form:

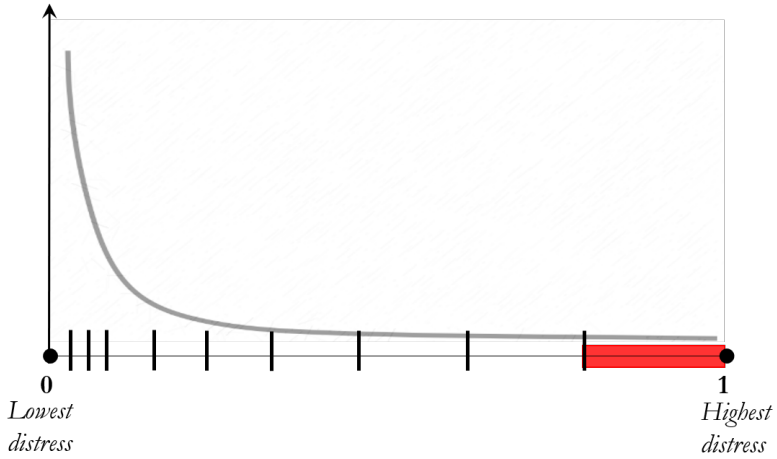
$$f(\mathbf{X}_{i,t-1}) = Pr(Y_{i,t} = 1 \mid \mathbf{X}_{i,t-1} = x) \tag{1.1}$$

where $Y_{i,t}$ is the binary realization of the outcome at time t that takes the value 1 if the ith firm exits the market and takes the value zero otherwise, while $\mathbf{X}_{i,t-1}$ is the P-dimensional vector of firm-level predictors in the previous time period, where P is the number of predictors included in the model. The functional form linking the predictors to the outcomes is determined by the generic supervised machine-learning procedure used to predict out-of-sample information. In short, the generic algorithm chooses the best in-sample loss-minimizing function in the form:

$$\arg\min \sum_{i=1}^{N} L(f(x_{i,t-1}), y_{i,t}) \quad over \quad f(\cdot) \in F \qquad s.t. \qquad R\big(f(\cdot)\big) \leq c \tag{1.2}$$

where F is a function class from where to pick $f(\cdot)$, and $R\big(f(\cdot)\big)$ is the generic regularizer that summarizes the complexity of $f(\cdot)$ (see also Mullainathan and Spiess, 2017). In our case, the function $f(\cdot)$ is an element from the family of classification trees or a combination of them. The set of regularizers, R's, will change according to the standards adopted by each method. Ultimately, each algorithm takes a loss function $L(f(x_i), y_i)$ as

Figure 2: Fictional distribution of failure's probability



input and searches for the function that minimizes the prediction losses.

To determine the region of highest financial distress, we show in the following analysis that a cutoff of 0.9 minimizes the combination of false-positive (false prediction of firm failures) and false-negative (false prediction of active firms) rates. This empirical result supports the choice of the highest decile of the predicted risk distribution as the optimal threshold for identifying cases of critical financial distress. We will also prove that using the highest decile to identify *zombie firms* is successful, as this is the segment where prediction accuracy is highest, and after which the probability of transitioning back to lower levels of financial distress is minimal.

We would like to emphasize that we are not interested in identifying the causes of firm failure, since we are dealing with a pure prediction problem,— i.e., the probability of the failure of firm i at time t. Nevertheless, we perform variable selection to evaluate the contribution of the predictors to the estimated risk. We show how the predictors of failure can change over time, under the circumstances in which we make the predictions, each time there is an update with new out-of-sample infor-

mation.

1.3.1 Decision tree learning for failure prediction

Consistent with the literature on bankruptcy and firm exit predictions presented earlier, we derive—given new out-of-sample information—a prediction for the failure of each firm based on its current financial accounts, both for established firms that have operated in previous periods and for firms entering the market for the first time. From another perspective, we interpret this probability range as the degree of risk of an investor who has no information other than that contained in the current financial accounts.

Our study introduces an innovative approach to failure prediction by demonstrating the remarkable effectiveness of tree-based machine-learning models, especially in the presence of missing data. While neural networks are widely considered the best option for prediction tasks with homogeneous data such as images and text, they often fail for "problems with heterogeneous features, noisy data, and complex dependencies" (Prokhorenkova et al., 2018). Decision tree–based algorithms are considered suitable tools in these cases due to their flexibility and high performance. The 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 shows how binary partitioning works in practice, using a simple example with only two predictors.

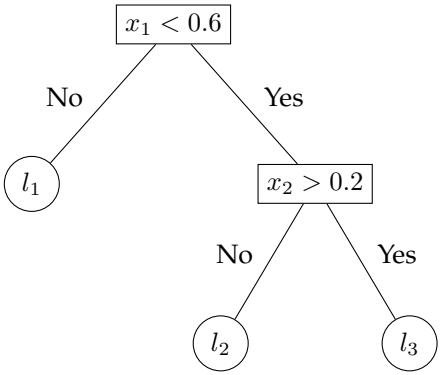
Ensemble methods have been developed to improve the stability of decision tree estimates, which combine multiple weak learners into one strong learner. Bagging and boosting are two popular methods for this purpose (Breiman, 1996; Freund, Schapire, et al., 1996). In line with recent developments in the literature, we perform a comparative evaluation of firm failure prediction using two state-of-the-art approaches: XGBoost (T. Chen and Guestrin, 2016) and BART-MIA (Chipman, George, McCulloch, et al., 2010). Both statistical models are specifically designed to deal with missing values, which makes them ideal for dealing with severe cases of missing financial data in corporate accounts. As illustrated

in Section 3.4, the absence of financial data does not follow random patterns, or at least does not follow the failure of a company completely randomly. When a company is financially distressed and/or smaller, it is more likely that data is missing from some accounts. Simply discarding the missing observations would introduce selection bias and exclude certain categories of firms with a higher probability of failure. Crucially, both XGBoost and BART-MIA include patterns of undisclosed accounts as a feature of the model.

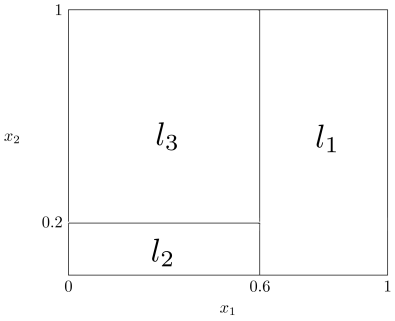
XGBoost is an ensemble method based on decision trees with a gradient boosting framework. It has proven to be highly competitive and reliable on various prediction tasks, mainly due to its efficient use of computation time and memory resources (Gumus and Kiran, 2017; Abbasi et al., 2019; M. Li, Fu, and D. Li, 2020). It is equipped with various optimization techniques and tools, such as a percentile-based split-finding algorithm, parallelized tree building, a depth-first approach to tree pruning, and efficient handling of missing values. XGBoost uses a technique called "Sparsity-aware Split Finding", also known as default directions (T. Chen and Guestrin, 2016) or block propagation (Josse et al., 2019), to deal with missing values for continuous variables. All observations reporting missing values are directed toward the branch with the least error magnitude. On the other hand, categorical variables require one-hot encoding before training, with missing values encoded as a separate category during preprocessing. In this way, the algorithm can make informed decisions based on the presence or absence of missing categorical data. In our application, the categorical variables have no missing values. Therefore, there is no need to encode missing values for them. Moreover, XGBoost includes regularization techniques to prevent overfitting (Bentéjac, Csörgő, and Martínez-Muñoz, 2021).

BART-MIA is a robust Bayesian ensemble of trees methodology that combines the traditional BART (Chipman, George, McCulloch, et al., 2010) with a variant called MIA (Twala, Jones, and Hand, 2008). This variant was specifically designed to deal with MNAR (Missing Not At Random) patterns in the data. BART is a well-established method that has shown excellent performance on various prediction tasks (Murray,

Figure 3: An example of binary tree



(a) Binary tree



(b) Feature space

Notes: 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.

2017; Linero and Y. Yang, 2018; Linero, 2018; Hernández et al., 2018). This can be attributed to two key factors: first, BART contains a noise component that mitigates the overfitting problem common to random forest methods, and second, it has shown consistently strong performance on standard model specifications, avoiding time-consuming hyperparameter tuning procedures. This property is particularly important because it reduces the researcher's dependence on parameter choice and minimizes the computational time and cost associated with cross-validation. See the appendix A.3 for the technical details of the two models.

1.3.2 Validation against other machine-learning techniques

We compare the predictive performance of our methods that account for missing values, XGBoost and BART-MIA, with the conditional inference tree (Hothorn, Hornik, and Zeileis, 2006), the random forest (Breiman, 2001), and the super learner (Van der Laan, Polley, and A. E. Hubbard, 2007). The conditional inference tree we use is a simple variant of the CART algorithm (Breiman et al., 1984), based on a significance testing procedure that avoids bias toward variables with many possible splits (see Odén, Wedel, et al., 1975; Loh, 2002; Hothorn, Hornik, and Zeileis, 2006). The random forest is an ensemble method that combines different trees to obtain stronger predictive power. Each tree is created by randomly selecting different variables from all possible predictors and randomly selecting a subset of the total number of observations (see also Breiman, 2001). The super learner (Van der Laan, Polley, and A. E. Hubbard, 2007) is based on a weighted combination of other algorithms. We use a non-negative least squares (NNLS) version of it, which ensures that the assigned weights are non-negative. This prevents negative contributions, which have no intuitive meaning in most contexts. Furthermore, in our variant of the NNLS version, the assigned weights are normalized so that they sum to 1. The normalization ensures that the final prediction is a weighted average of the base learners' predictions, with each weight representing the relative importance of the base learner in the ensemble. We build it as a convex combination of the following models: logistic regression, CART, random forest, BART, and XGBoost.

To evaluate the performance of the models, we use a standard five-fold cross-validation for panel data, where the dataset is divided into five groups and the model is trained on four of them, while the fifth group is used as a validation set, iteratively repeating this process. In this way, we can comprehensively evaluate the predictive performance of the model over the entire time series.

To ensure comparability with the other models that do not consider missing values, we perform a complete-case (CC) analysis. The latter is a widely used approach for dealing with missing values in models that are not designed to do so. In this approach, all observations with at least one missing value in the predictors are discarded. Although XGBoost and the BART-MIA models could use the entire dataset without excluding missing observations, we perform a missing-aware analysis to ensure fair performance evaluation of all models: we train and test XGBoost and BART-MIA using fivefold cross-validation with identical dimensions as the original folds but allow missing observations to be included in the sample. For clarity, we use the terms XGBoost and BART-MIA for the models trained on observations with missing values. In this case, XGBoost uses default directions to deal with missing values, while BART-MIA uses the MIA procedure. On the other hand, we will simply refer to the models trained on CC data as CC-XGBoost and BART.

Table 3: Models' horse race: performance measures

(a) CC 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

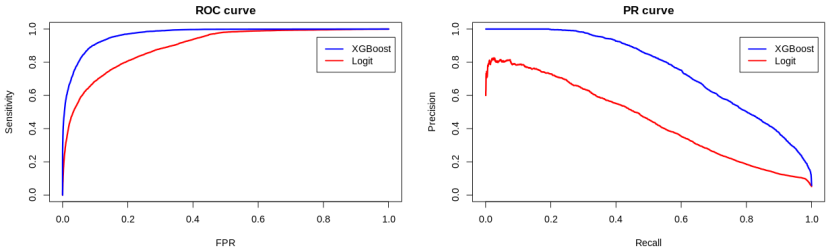
(b) Missing-aware analysis

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

Notes: 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.

Table 3 shows the results of the horse race of the models. It is clear that the missing-aware models consistently outperform the state-of-the-art methods involved in the CC analysis. To compare the predictive power of the different methods, we show five different performance measures commonly used for classification problems: the area under the receiver operating characteristic (ROC) curve (AUC), the area under the PR curve, F1-score, Balanced ACCuracy (BACC), and adjusted R^2. Both AUC and PR vary between 0 and 1, with 0 indicating complete misclassification and 1 indicating perfect prediction. The AUC (Hanley and McNeil, 1982) is a general measure of predictive power that tells us the extent to which we are able to classify failures *vis à vis* non-failures, hence with an accent on the *false discovery rate* (FDR). PR is particularly useful for our scope because it takes into account both the total proportion of true failures that we are able to predict in the data (i.e., the *sensitiv-*

Figure 4: Goodness-of-fit scores



Notes: 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.

ity/recall of the predictions) and for the proportion of predicted failures that turn out to be true failures (i.e., the *precision* of the predictions). Indeed, assessing *sensitivity/recall* alone could be misleading in a zero-inflation environment such as ours, where the number of non-failures systematically exceeds the number of failures.[10] The F1-Score (Van Rijsbergen, 1979) and the BACC (Brodersen et al., 2010) are used for cases of unbalanced data. The former is based on a harmonic mean of *precision* and *recall*, while the latter is a simple average between the rate of true positives and the rate of true negatives from our predictions.

In the CC analysis, the best-performing model is the super learner, which, however, stops at 0.9231 and 0.5463 in terms of AUC and PR, respectively.[11] The results of the missing-aware analysis differ significantly from those of the CC ones. XGboost shows a PR of 0.7591 and an AUC of 0.9685, followed by BART-MIA, whose performances are almost identical. However, training the BART-MIA model is computationally intensive and takes 45 times longer than XGBoost. Due to its signifi-

[10] For more details, see Saito and Rehmsmeier, 2015; Fawcett, 2006.

[11] It is noteworthy that the Super Learner performs better than the Random Forest in terms of accuracy, but not in terms of F1-Score. The reason is that the Super Learner algorithm (Van der Laan, Polley, and A. E. Hubbard, 2007) is optimized to find a convex combination of algorithms that minimizes the accuracy of the ensemble method. The latter strategy is not optimal in our case because the dataset is unbalanced.

Table 4: Goodness-of-fit: *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

Notes: For *DtD*, *Z-Scores*, a firm is predicted to fail if its score falls below a certain threshold. The thresholds for classification are determined by the first 10 percentiles of their in-sample score distribution. In contrast, XGBoost predicts failure when the score is above a certain threshold, and classification cutoffs are determined by the top 10 percentiles of the own in-sample score distribution.

cantly lower computational cost and higher predictive power, we choose XGBoost as the optimal prediction algorithm for the following analysis.

Interestingly, the most notable improvement comes from higher precision. Figure 4 shows the average ROC and PR curves over the fivefolds for CC logit and XGBoost to illustrate the difference in performance between the traditional approach and our best-performing missing-aware machine-learning method. This is also reflected in the larger departure of the missing-aware models with respect to the performance measures using precision (namely PR and F1-score) reported in Table 4. This is critical to our objective as it means that missing-aware models have better predictive power in identifying companies that will close within a year based on the predictions.

To ensure the robustness of our results, we perform a number of robustness checks in Appendix A.5. We investigate whether (i) imputing missing observations could improve the performance of predictive

models that do not include information on missing attributes and (ii) excluding liquidations from the set of firm failures would affect the performance of the model. We find that the performance of traditional models increases significantly when missing values are considered, especially for tree-based methods. This result is consistent with our observation that missing entries signalled by a constant value either over the entire dataset (out-of-range imputation) or over a specific feature (median imputation) are included in tree splitting. As a result, the model can now account for missingness in a similar way to the standard directions and MIA techniques, leading to comparable results. With respect to (ii), we test whether predictive performance depends on the type of failure (e.g., bankruptcy, dissolution, merger, acquisition, and liquidation). We find that excluding the largest group of business failures—i.e., liquidations—has little effect on the performance of the models, suggesting that the dependence on the type of failure is negligible.

1.3.3 Validation against proxy models of credit scoring

So far, we have compared the predictions of failures from different econometric and machine-learning techniques and concluded that XGBoost is the best choice in terms of predictive power and computational burden. Here, we compare our baseline predictions with widely known proxies for firm-level credit scores: the *Z-Scores* (Altman, 1968) and the *DtD* (Merton, 1974). *Z-Scores* involve putting a selection of financial ratios (profitability, leverage, liquidity, solvency, and volumes of activity) into an equation with some weights to proxy their relative importance. The weights are taken from the literature and from previous scholarly estimates of the relative importance of these indicators in assessing a firm's distress. In this way, a threshold is obtained, the crossing of which indicates a high probability of future bankruptcy. Unlike *Z-Scores*, the *DtD* by Merton (1974) focuses specifically on a firm's ability to meet its financial obligations. The original intuition is that a firm's equity can be modelled as a call option on its assets. Thus, to build such a model, one must combine firm-level accounts (firm assets, debt, and market value) and information from financial markets (risk-free interest rate and stan-

dard deviation of stock returns). Finally, one puts the variables into an equation that gives the value of a theoretically fair call option.[12]

We evaluate the predictive power of both *Z-Scores* and DtD. The precision and false discovery rate (FDR) are given in Figure 4. Similar to our previous analysis, we performed fivefold crossvalidation. We start with the non-missing observations in *Z-Scores* and *DtD* and randomly split them into fivefolds. Based on the *Z-Scores* and *DtD* measures, distressed firms have lower scores. Therefore, we used the first 10 percentiles of the in-sample scores as cutoffs for classifying the out-of-sample observations. We found that the *DtD* predictions have higher precision (0.3314 vs. 0.2239) and lower FDR (0.6686 vs. 0.7761) than the *Z-Scores*. We then train an XGBoost model with the same cross-validation routine and use the in-sample percentiles as cutoffs to classify the out-of-sample observations. The XGBoost model considers the highest-risk firms to be in the right tail, unlike *Z-Scores* and *DtD*. Therefore, the first column of Table 4 represents the bottom ten percentiles (1-10) for *DtD* and *Z-Scores*, while for XGboost the top ten percentiles (99-90) symmetrically. It is clear that XGboost outperforms both *DtD* and *Z-Scores* in all percentiles.

1.3.4 Unboxing the black box: Shapley values for predictors

In the previous sections, we validated our proposed XGBoost model against state-of-the-art financial indicators, traditional econometric models, and widely used machine-learning techniques. Despite its excellent predictive performance, the nonlinear relationships learned by XGBoost are not directly observable, leading to the recurring criticism that machine-learning models are black boxes despite their improved accuracy. However, there are a number of tools to shed light on complex nonlinear relationships in machine-learning predictions to make models interpretable (Lundberg et al., 2020).[13] Popular methods for assessing

[12]Following the insights of the distance-to-default model, Black and Scholes, 1973 developed their widely known model based on the observation that one can eliminate a systemic risk component by hedging an option.

[13]Interpretability is a non-mathematical concept, but is often defined as the degree to which a human can understand the cause of a decision or consistently predict the outcomes

predictor importance include permutation importance or Gini importance for tree-based models (Breiman, 2001), LIME (Ribeiro, Singh, and Guestrin, 2016), DeepLIFT (Shrikumar, Greenside, and Kundaje, 2017), and Shapley values (Strumbelj and Kononenko, 2010).

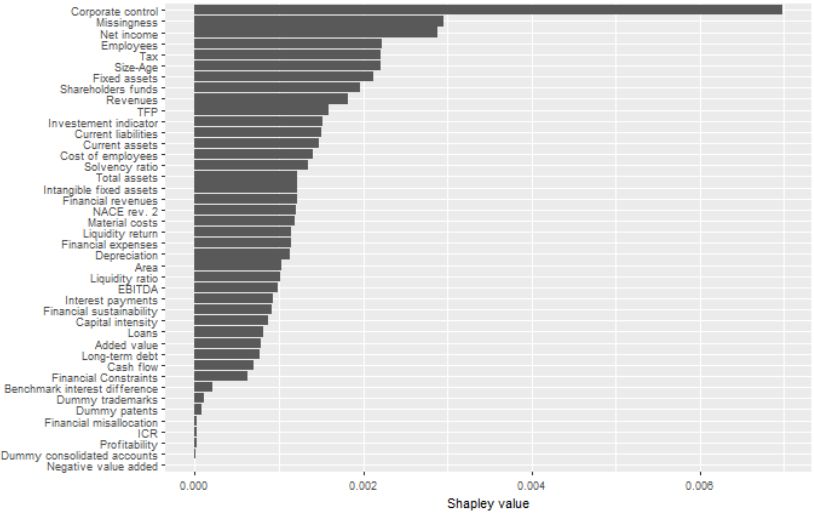
We choose to use Shapley values because they are able to not only reveal the complex patterns linking the predictors and the outcome but also guarantee a number of desirable properties: *efficiency, null payoffs, symmetry, monotonicity* and *linearity* (Rozemberczki et al., 2022). In the case of a general ML model, these properties state that: the importance of individual variables must add up to the goodness-of-fit of the model trained with the entire set of variables (*efficiency*)[14]; a variable that does not improve the goodness-of-fit of the model is assigned a contribution of zero (*null payoff*); two variables that make the same marginal contribution to the global goodness-of-fit have the same importance (*symmetry*); if variable A consistently contributes more to the global goodness-of-fit of the model than variable B, the Shapley value of A must be higher than that of B (*monotonicity*); for two subsets of the same data set, the global Shapley value of a variable is equal to the sum of the Shapley values calculated separately for the same variable in the two subsets (*linearity*).

An intuitive way to understand how Shapley values for variable importance work is to include each variable in the predictive model in random order. All variables in the model contribute to the final predictions (and thus to the goodness-of-fit of the model). The Shapley value of a variable is the average change in prediction (and goodness-of-fit) experienced by the set of variables already in the model when that variable is included (Molnar, 2020). From a mathematical perspective, 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

of the model see, e.g., Kim, Khanna, and Koyejo, 2016; T. Miller, 2019; K. Lee, Falco J Bargagli-Stoffi, and Dominici, 2020.

[14]This property is crucial because it allows quantifying the contribution of each variable to the overall performance of the model.

Figure 5: Shapley values for the variables in the predictive model



generic variable m is:

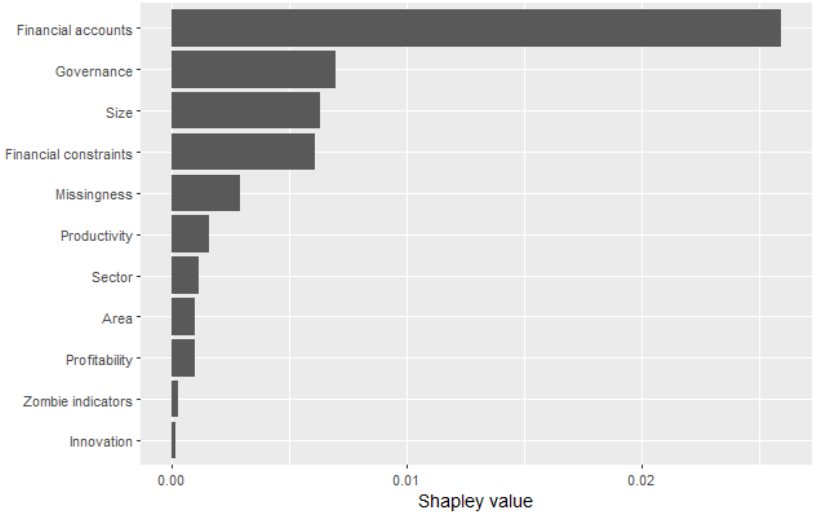
$$\phi_m(v) = \frac{1}{q} \sum_{S \subseteq \{1,\ldots,q\} \setminus \{m\}} [v(S \cup \{m\}) - v(S)] \frac{|S|!(q - |S| - 1)!}{q!}. \quad (1.3)$$

Using equation (1.3), 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.

The results of our Shapley value analyses are shown in Figures 5 and 6. Figure 5 reports the results of the average Shapley Values over the time frame of our analysis for each variable used as input. Figure 6 shows the same value, but this time all variables are aggregated into some economically relevant groups.[15]

[15] A full description of how variables are aggregated into groups is described in Appendix Table A1. Governance variables are indicated by (G), financial constraints by (FC), financial accounts by (FA), zombie indicators by (ZI), area by (A), innovation by (I), productivity by (PDC), sector by (SE), profitability by (PFT), size by (SI).

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



First, we note that no single financial indicator has better predictive power than the aggregate of indicators. On their own, most indicators have relatively low predictive power. The financial indicators with higher Shapley values—i.e., that contribute most to the model's predictions—are the corporate control indicator(*Corporate control*), the missingness in the financial accounts (*Missingness*), some profit and loss characteristics (*Net income, Taxes, Revenues*), the size and age indicators (*Employees, Size-age*), the resources invested in the company (*Shareholders funds*) and the long-term physical assets (*Fixed assets*).[16] When analyzing the groups of indicators, it appears that original financial accounts with no further elaboration carry the greatest explanatory power. This result is reasonable considering that most predictors fall into this category. In addition, the governance, size, and financial constraint indicators also have relatively high predictive power. Although they compete with groups consisting of multiple predictors, missingness remains an important factor. On the other hand, firms' productivity, sector, and geographic loca-

[16]See Table A.1.1 for a full description of the predictors and their construction.

tion of firms also have high predictive power but are of secondary importance compared to the aforementioned indicators.

Finally, to check for robustness, we conduct a logit-Least Absolute Shrinkage and Selection Operator (LASSO) analysis to evaluate, using a shrinkage method, the most frequently selected variables that predict a firm's distress. We find substantial overlap, with corporate control, profit and loss characteristics, and firm size-age indicators among the most important predictors. The full analysis can be found in Appendix A.4.

We conclude this section with a discussion of the possibility that firms manipulate financial accounts and the predictive power of missing values in this case. We argue that there are two reasons for the occurrence of missing values in firms' financial accounts:

1. Smaller firms are often exempt from reporting complete information to national registries if they fall below certain size thresholds because it may be too costly for them to implement complex accounting procedures.[17]

2. Some financial accounts are optional each year or may be incomplete (e.g., number of employees, management of inventories, cash prospects, etc.).

In any case, there is room for manipulation by firms that know that banks and policymakers do not routinely use missing values to predict their financial distress. They have two choices: they can provide truthful information and thereby disclose their true financial distress or they can provide false information and thereby potentially commit accounting fraud. In the first case, the missing values are no longer meaningful, but our algorithm can still rely on the disclosed financial accounts for prediction. In the second case, we would have false negatives, i.e., companies that are predicted to be financially viable but are not. More generally, we conclude that a limitation of our approach is that we assume that companies do not report false information; otherwise, we would have statistical noise that reduces prediction accuracy.

[17] In footnote 7 of the manuscript, we indicate the combination of size thresholds that the Italian law provides.

1.4 A case for *zombie firms*

In our analysis, we showed how we can use machine-learning to predict firms' failures. But how do we spot a *zombie firm*? There is no consensus on the exact meaning of the *zombie* status, other than its suggestive power. To date, scholars have merely adopted various thresholds based on a proxy assessment of one or more available financial indicators in the absence of more precise theoretical guidance. Ideally, a company's competitiveness and financial constraints should be viewed from a dynamic perspective that considers the entire horizon of future events. Considering all future threats and opportunities and their impact on profit and net cash flow at the firm level could be the theoretical equivalent of "*zombie firms*". Without the latter, empirical identification is left to the creativity of academics and practitioners. Caballero, Hoshi, and Kashyap (2008) define *zombies* as firms that receive subsidies in the form of bank loans after observing how interest payments compare to an estimated benchmark of debt structure and market interest rates. Müge Adalet McGowan, Andrews, and Millot (2018) assume that *zombies* are old firms that have persistent problems meeting their interest payments, although they focus a policy discussion on the macroeconomic impact of low-productivity firms. Bank of Korea (2013) explicitly examines when the ICR is below one over three years. Bank of England (2013) disregards financial management and considers firms that have both negative profits and negative value added, thus focusing on a firm's core activity.

In our view, the direction of the work so far is clear: scholars and practitioners want to infer companies' future viability from their current financial accounts. If they do not appear to be in good shape, it is likely that the company will be in trouble in the near future. From this perspective, the empirical classification of *zombie firms* is a perfect case study for applying machine-learning techniques to firm-level data. It is a call to use in-sample information to predict an out-of-sample event.

Strengthened by the previous intuition, we propose an identification of *zombies*, starting from the predictions of failure made in Section 1.3. We propose to classify as *zombies* those firms that are at the right end of the

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

risk distribution and for which the chances of recovering from financial distress are minimal for at least 3 years. In notation, we first consider the deciles along the predictions $f(\mathbf{X}_{i,t-1})$, where each $q_{j,t}$ is the threshold for the jth decile of the probability of default at time t:

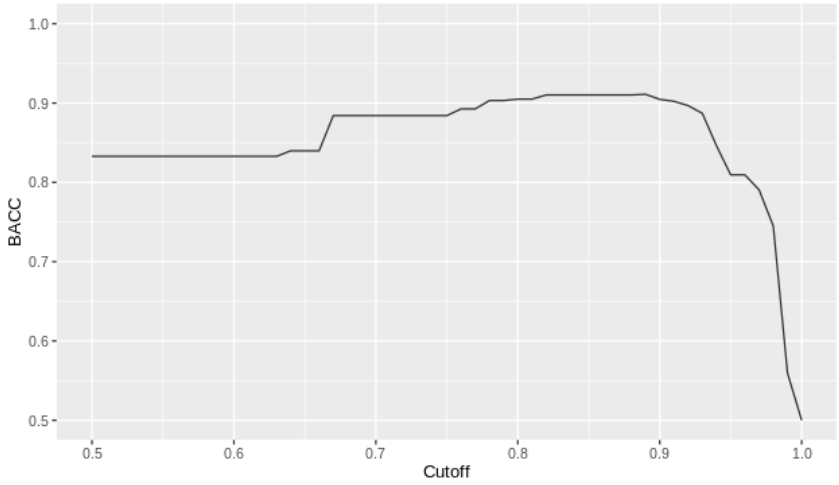
$$Q_{j,t} = \begin{cases} 1 & \text{if } f(\mathbf{X}_{i,t-1}) \geq q_{j,t} \cap Y_{i,t} \neq 1, \\ 0 & \text{otherwise.} \end{cases}$$

where $Y_{i,t} \neq 1$ indicates that the ith firm did not fail yet at time t.

In Table 5, we report a transition matrix for the firms that did not fail, based on elaborations over the entire period of analyses (2008-2017). We find that a significant fraction of the firms that our predictions place beyond the 9th decile in a representative year t do not have much chance to improve in t + 1. In fact, most of them (46%) remain stuck in the same highest-risk category, and only 12% are able to recover and reach a more reasonable level of financial distress, i.e., below the 6th decile. Interestingly, the 9th decile is difficult to reach from the bottom of the distribution, as only 11%, 7%, and 3% of firms from the lower deciles, respectively, reach a situation of highest distress. Moreover, around 78% of the companies remain below the 6th decile throughout the analysis period.

In general, we can say that, according to the information we have observed, a viable firm does not easily shift into financial distress, but if it does, it is difficult to recover from it. With this in mind, 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. We accomplish this task by maximizing the BACC since it corre-

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

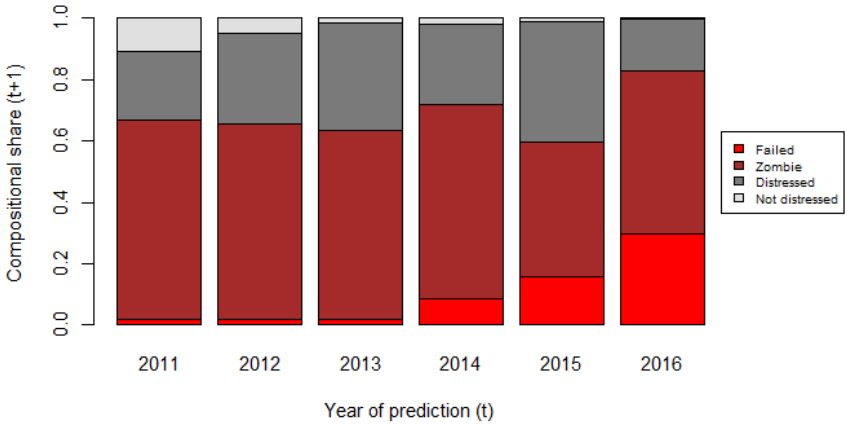


sponds to a convex combination of the true positive and negative rates, which in turn are complementary to the false positive and negative rates. Moreover, the concept of zombie firms is closely related to the concept of false positives since these firms are expected to fail but remain active. Therefore, focusing on the false positive and negative rates within this framework is natural. According to our results in Figure 7, the BACC peaks at about 0.9.

The latter result provides a first justification for using the 9th decile as a cutoff. Nevertheless, we want to give firms a trial period to discount the break-even strategies of some firms in the short run, e.g., in the case of newly formed firms and start-ups. For all the above reasons, we define *zombie firms* those firms that *persist* at least three years beyond the 9th decile of the risk distribution:

$$\mathbb{1}(\sum_{t=1}^{3} Q_{j,t} = 3) \tag{1.4}$$

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



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

where the risk distributions are estimated with the XGboost, as proposed in Section 1.3.

Finally, Figure 8 shows a transition analysis of zombie firms based on our definition from Equation (1.4).We calculate the fraction of *zombies* that fail, remain in zombie status, transition to a lower (but still significant) risk category, or transition to a non-distressed state in the following year. These results suggest that the most common outcome is remaining in zombie status, which applies to nearly 50% of them over multiple years. In addition, a substantial proportion of zombies transition to a lower-risk category. However, only a small fraction of them completely deviate from their current state, resulting in either failure or absence of distress.

1.4.1 *Zombies* in Italy

In this section, we give some coordinates for the phenomenon of *zombies* in Italy. First, we provide an overview of their evolution along the macroeconomic cycle. Then, we show how they differ from the rest of the viable firms that survive the market. Finally, we provide a geographical and cross-sector comparison, taking into account their financial performance and their ability to generate added value.

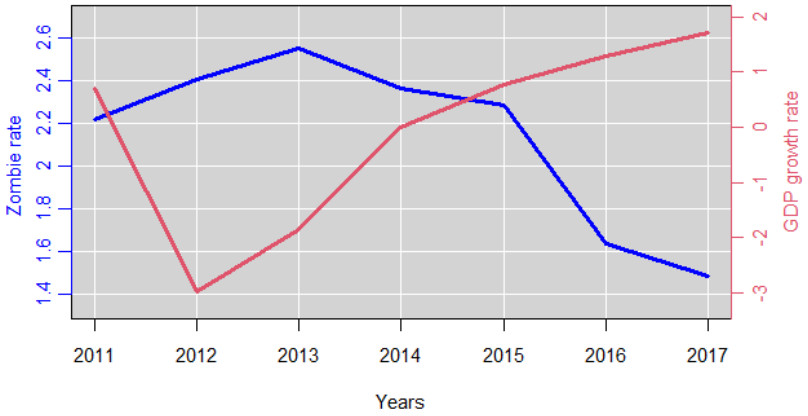
In Figure 9, we plot the share of *zombies* in our analysis period against the Gross Domestic Product (GDP) growth rates observed over the same period. Interestingly, we find that a range between 1.48% and 2.55% of manufacturing firms are on the verge of bankruptcy. The share of *zombies* is higher immediately after the financial crisis in 2011 and then decreased from 2013. In fact, the presence of *zombies* seems to be related to the business cycle. The latter is an interesting finding for further analysis but beyond the scope of this paper. We presume it makes sense for *zombie firms* to be countercyclical: many firms can be pushed to the brink of bankruptcy in times of crisis, while a few financially distressed firms can find a way to recover when the recovery begins.

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

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

Notes: The table shows the coefficients of the linear models for panel data with fixed effects at the region and industry levels. We use pooled Ordinary Least Squares (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 1 if the firm is classified as *zombie* in a given year of our sample, and 0 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. Statistical significance at the 1% level is indicated by ***.

Figure 9: *Zombie firms* and the economic cycle



Notes: 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.

To fully capture their economic importance, we briefly show in Table 6 what distinguishes *zombies* from the other firms. Controlling for industry and region, we find that zombies are on average less productive, i.e., 19.7% and 45%, respectively, in terms of TFP and labor productivity. Healthy firms are also consistently larger on average, selling about 200% more and employing 17% more workers than zombie firms in the same sector and region. Our results are consistent with the idea that these firms drag down aggregate productivity while hindering the redistribution of resources (Andrews, Muge Adalet McGowan, Millot, et al., 2017; Müge Adalet McGowan, Andrews, and Millot, 2018).

Eventually, we lay out how the segment of zombies that we find according to our working definition compares to the segments of firms that:

1. have problems in making their interest payments because their Interest Coverage Ratio (ICR);

2. have problems in their core economic activity because they are de-

stroying value, i.e., their value added is negative.

Interestingly, both ICR and negative value added have occasionally been proposed as indicators of *zombieness* in previous literature (see Appendix A.1 for more details on these indicators and their previous use).

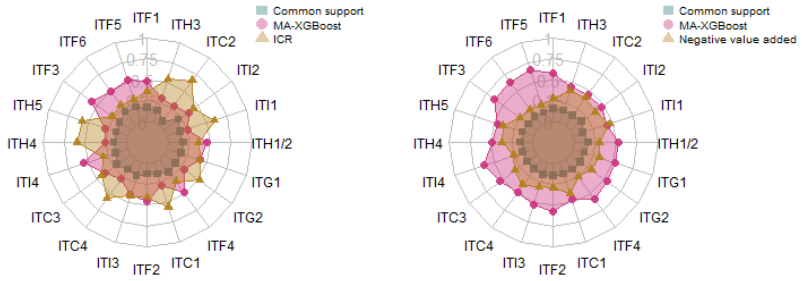
In Figure 10, the rays of the radar show for NUTS 2-digit regions the proportions of *zombie firms* that we detect with XGBoost (circles) compared to firms that have an ICR of less than 1 (triangles in panel a) and firms that experience negative value added (triangles in panel b). We find that there is indeed common support when firms are financially distressed because their ICR is too low, and at the same time they are also classified as *zombies* according to our algorithm. This is evident from a common core region in the radar graph bounded by square nodes. However, the two segments do not coincide. If we consider only the ICR, we may have zombies that are obviously not in distress but are classified as such by our algorithm. It is noticeable that this occurs mostly in southern and central Italy: Abruzzo (ITF1), Basilicata (ITF5), Calabria (ITF6), Campania (ITF3), Lazio (ITI4), Molise (ITF2), and Puglia (ITF4) host a relatively large proportion of zombies according to XGBoost.[18] In these cases, the cause of distress may lie in other aspects of business activity, as our methodology allows us to use information from a wide range of financial predictors covering different aspects of firms' economic life.

At the same time, Figure 10a also identifies firms that have liquidity problems because their ICR is less than 1, but our algorithm does not classify them as *zombies*, possibly because they are still solvent thanks to profitable economic activity that allows them to eventually meet their financial obligations and thus overcome temporary liquidity constraints shortages. This is particularly evident in North and Insular Italy, including Emilia-Romagna (ITH5), Friuli-Venezia Giulia (ITH4), Lombardy (ITC4), Piemonte (ITC1), Sardinia (ITG2), Tuscany (ITI1), Valle d'Aosta (ITC2), and Veneto (ITH3).

On the other hand, in Figure 10b, we show that a relatively small

[18]Not surprisingly, patterns of missing values are also more pronounced in southern and central Italy. Between 2008 and 2016, ICR values are missing for 81.79% of firms in central Italy and 76.24% of firms in southern Italy on average

Figure 10: *Zombie firms* and geography



(a) *Zombies* and interest payments

(b) *Zombies* and value destruction

Notes: The rays of the radar show, at the regional level (NUTS 2-digit), the proportion of *zombie* firms vs. firms that have an ICR of less than one (panel a) and vs. 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 1. Panel (a) and (b) include a total of 30,380 and 24,351 observations, respectively. See Appendix Table A.1.3 for the legend of the NUTS 2-digit codes.

segment of firms destroy economic value because what they sell has less value than what they buy as inputs. They are problematic; in fact, most of them are classified as *zombies*. However, there is a significant portion of zombies that show positive value added but are still struggling financially and are therefore detected as *zombies* by our algorithm. This scenario extends over the entire country but peaks in some central and southern regions: Abruzzo (ITF1), Basilicata (ITF5), Calabria (ITF6), Campania (ITF3), Lazio (ITI4), Molise (ITF2), and Puglia (ITF4). The descriptive statistics in Figure A.2.2 confirm our intuition that only a comprehensive battery of predictors, including missing values, can provide a clear picture of the relevance of the phenomenon of zombie firms. Finally, Appendix Table A.1.4 and Figure A.2.2 provide a snapshot of the proportion of zombies across industries (2-digit NACE Rev.

2). As largely expected, firms that have problems with interest payments ($ICR < 1$) are always a larger segment than firms that we identify as *zombies*. On the other hand, the share of firms with negative value added is generally smaller than the share of *zombie firms*.

1.5 Discussion

In this contribution, we show how statistical learning can infer non-trivial information on a set of financial indicators, and we successfully classify firms into risk categories after training on past failures. Our preferred algorithm is XGBoost, which outperforms other well-known econometric and machine-learning methods, especially in the presence of missing patterns of financial accounts.

Thanks to our machine-learning approach, we can also reduce prediction errors compared to traditional credit scoring tools such as the *Z-scores* and the *DtD*. Therefore, we propose to classify *zombie firms* as those firms that remain in high-risk status because they are at the right end of the distribution of our predictions for at least 3 years, beyond the 9th decile of risk, where we find that the chances of recovery to a smaller risk of distress are minimal.

The preceding evidence suggests that identifying *zombies* may be crucial for financial institutions to avoid wasting credit resources on insolvent firms when the latter survive the market only thanks to some adverse selection mechanisms that arise from imperfect financial markets. However, from a more general perspective, we believe that our exercise can be useful in identifying a part of an economy that is in trouble. The issue is all the more critical because recent studies discuss how *zombies* can hinder the growth potential of many countries by impeding the redistribution of productive resources in modern economies (Andrews, Muge Adalet McGowan, Millot, et al., 2017; Banerjee and Hofmann, 2018; Andrews and Petroulakis, 2019).

Indeed, we find that Italian *zombies* have lower levels of TFP and are mainly found in smaller firms. Interestingly, we note a possible relationship with the business cycle that would be worth investigating further in

the future: the share of *zombies* was higher after the two financial crises in 2008 and 2011 but then declined as the recovery took hold.

It is beyond the scope of this work to make predictions for the post-pandemic scenario. However, we anticipate that the problem of separating companies that can stand on their own two feet from those that hide their inability to pay will become all the more urgent now that policy-makers have finally withdrawn financial support programs. The challenge is to avoid further misallocation of resources, which would slow down the much-needed economic recovery.

Chapter 2

Founding Conditions, Learning and the Economic Success of Young Firms

The chapter is co-authored with Mark Kattenberg, Massimo Riccaboni, and Erik Stam.

Disclaimer: AI has been employed in the preparation of this chapter, specifically for grammar and language refinement. All content, ideas, and interpretations remain the original work of the author and coauthors.

2.1 Introduction

Many high-income economies are facing declining productivity growth. In order to boost productivity growth, a process of creative destruction is required, in which new highly productive firms displace less productive firms and increase overall productivity (Bartelsman and Doms, 2000; Aghion, Akcigit, and Howitt, 2014; Acemoglu et al., 2018). This process of creative destruction does not take place to a sufficient extent because there are too few new, highly productive companies (Akcigit and Ates, 2021; Decker et al., 2016). This may be due to a lack of entry in general

or a lack of entrants achieving high productivity. An extensive literature has been developed on the survival of new firms (Audretsch and Mahmood, 1995; Geroski, José Mata, and Portugal, 2010; Manjón-Antolín and Arauzo-Carod, 2008; Soto-Simeone, Sirén, and Antretter, 2020), but surprisingly very little is known about how productive firms emerge. This paper aims to improve our understanding of the survival and economic success of young firms. To achieve this, we present two interpretations of the survival and economic success of young firms inspired by two theories of firm dynamics: noisy learning (Jovanovic, 1982) and capability learning (Teece, Pisano, and Shuen, 1997).

The theory of noisy learning states that firms learn about their productivity and market fit through experience. Some new firms discover that they have a viable business model and grow, while others exit early on due to a lack of product-market fit and resulting profitability. Importantly, in this framework, the initial conditions determine the survival and long-term success of the company. In contrast, capability learning theory posits that firms actively improve their resources, knowledge, and strategies over time. Adaptive firms that are able to develop new capabilities are more likely to be successful in the long term, even if their initial conditions are less favorable.

To test these theories, we focus on the Netherlands, a country with high entry rates but stagnating overall productivity. Several studies have analyzed young firm survival and success (cf. Geroski, José Mata, and Portugal, 2010; Fuertes-Callén, Cuellar-Fernández, and Serrano-Cinca, 2022, and for a recent review: Soto-Simeone, Sirén, and Antretter, 2020). Although many studies rely on commercial data sources, such as Bureau van Dijk—which often do not cover the full population of firms and lack detailed, year-by-year information on each entry cohort—some recent contributions have addressed these limitations by using administrative data. For instance, Fairlie et al. (2025) leverage the U.S. universe of startups to study the impact of entry policies on entrepreneurial outcomes, Cascarano, Natoli, and Petrella (2025) examine firm dynamics in Italy using matched employer-employee data covering the entire population of Italian startups. Similarly, Grazzi and Moschella (2018) examine the role

of small firms in productivity growth by drawing on a rich administrative dataset that links the universe of Italian firms with detailed export transaction records. However, despite these valuable efforts, the use of high-dimensional administrative data to track the full population of new firms over several years of their early development remains relatively limited. This study seeks to contribute to this emerging area of research.

This study analyzes comprehensive administrative data on the population of newly founded firms in the Netherlands from 2006 to 2021 and combines this data with a novel machine-learning methodology inspired by the seminal contribution of Mueller and Spinnewijn (2023). This approach makes it possible to quantify the persistence of firm success probability types over time and to assess the relative importance of dynamic selection versus within-firm learning in explaining economic success. The central research question of this study is: to what extent is the economic success of young firms determined by dynamic selection and within-firm learning? To answer this question, the study investigates whether the survival of young firms depends on the characteristics of the firm at the time of founding, and the changes in firm characteristics over time. It also examines whether the survival mechanisms of productive firms differ from those of unproductive firms.

The study constructs a high-dimensional dataset by bringing together over a hundred administrative sources from Statistics Netherlands. This dataset contains firm-level financial statements, employment data and ownership characteristics and allows a detailed analysis of how initial conditions and within-firm dynamics shape firm performance. An important contribution is the application of state-of-the-art supervised learning methods in combination with Shapley value decomposition and accumulated local effects analysis. These techniques enable the prediction of economic performance based on early-life course firm characteristics which allows us to quantify the heterogeneity in firm economic success probability types and its dynamics over time as firms age and gain experience.

Our results contribute to the growing literature on the determinants of firm success. First, we find that early-life firm characteristics are highly

predictive of long-term success—not necessarily due to direct causality, but because they are correlated with firm attributes that emerge later in the life cycle. This persistence in success probability types is observable both across the firm's life course and over time, and we interpret it as evidence in support of noisy learning. Moreover, our findings align with entrepreneurship literature, which highlights how the founding team affects a firm's initial conditions and, consequently, its long-term performance. In particular, Zhao, Song, and Storm (2013) and Jin et al. (2017) highlight the significant impact of founding team capabilities and composition on new venture performance, mediated by strategic positional advantages.

Our study is also related to the industrial dynamics literature. Cavallari, S. Romano, and Naticchioni (2021) show how firms born during recessions start on a larger scale and keep larger compared to those born during expansions, suggesting that the economic conditions at the time of entry have long-term implications for firm performance. Our findings strongly align with these findings and indicate that annual fluctuations in the success rates of young firms at the macro level appear to result from variations in the timing and decision to enter the market, rather than variations in the decision of how to enter the market.

The results of this study have significant implications for policy. If initial firm conditions largely determine long-term outcomes, policies to improve conditions for early-stage businesses, such as better access to finance, business mentoring, and managerial training, could have a lasting impact on economic productivity. Furthermore, the findings challenge traditional views on capability learning, and suggest that while firms may adapt over time, in most cases these adaptations do not fundamentally alter the long-term probabilities of success.

The rest of the paper is structured as follows. Section 2.2 provides an overview of existing theories of young firm survival and economic success. Section 2.4 describes the construction and content of our dataset. Section 2.3 presents the methodology. In particular, it presents a conceptual framework that relates our empirical analysis to noisy learning and capability learning. Section 2.5 presents our results on the predictive per-

formance and persistence of firms' success probability types. Section 2.6 derives the main implications for different types of learning. Finally, in Section 2.7, we draw conclusions.

2.2 Theories of young firm survival and economic success

In our work, we use two theories of firm dynamics to interpret the causes of the survival and economic success of young firms: the theory of noisy learning and the theory of capability learning. A firm is an entity with resources and capabilities that are used to produce and sell goods and services in specific markets. Moreover, a firm is characterized by its business model, which determines how it wants to make profits, what products it wants to sell, which market it targets, what investments are to be expected, and what the competition looks like.

2.2.1 Noisy learning

In the theory of noisy learning, firm dynamics depend on the learning process through which firms discover their productivity. Founders do not know the (potential) productivity of their firm in advance, and market interactions are used to obtain this information. Each firm enters the market with a specific configuration of resources and products (business model). The theory assumes that a firm type is given from the foundation, remains unobservable, and does not undergo significant adjustments over time. Those firms that are productive, grow and survive, while those that are not productive exit. This theory predicts that the survival rate of a company increases with both the current size (conditional on age) as well as with firm age, as larger firms tend to be more productive and companies that survive longer are more likely to have discovered that they have a viable business model. These predictions are consistent with the evidence from early empirical studies in industrial economics (see, among others, Evans, 1987b; B. H. Hall, 1986). Empirical studies have shown that the probability of survival increases with

the age and size of the firm (Dunne, Roberts, and Samuelson, 1989; Audretsch and Mahmood, 1994; Jose Mata and Portugal, 1994; Mitchell, 1994; Haveman, 1995; Sharma and Kesner, 1996; Buldyrev et al., 2020) albeit at a decreasing rate. In addition, studies that have focused on the post-entry performance of new firms have found that the probability of survival increases with the size of the new firm at the time of entry (Agarwal and Audretsch, 2001). Further cross-country evidence supports the idea that firm growth dynamics differ substantially across contexts, with important implications for long-run survival and productivity patterns (Hsieh and Klenow, 2014). Firm size matters not only at the beginning but also in relation to the growth of the company. The fact that a firm has grown in the past indicates that it has performed well and wants to become larger than it currently is. Consequently, it should have a lower probability of exit than its current size alone would suggest (José Mata, Portugal, and Guimaraes, 1995). Coad et al. (2013) integrates these findings and shows that the survival of a new firm depends on the resources available at start-up or accumulated through post-entry growth. A refined version of this noisy learning theory states that firms must reach a minimum efficient size to survive and that firms below this minimum efficient size have a lower probability of survival than firms beyond this minimum efficient size. Beyond this minimum efficient size, there may be no positive effect on survival, as shown by the decreasing effect of firm size on firm survival (Farinas and L. Moreno, 2000).

2.2.2 Capability learning

In the theory of capability learning, some new firms are better able to improve their resources and develop new products and markets than others, which makes these "adaptive" firms more likely to be successful. A key property of such adaptive firms are dynamic capabilities, which reflect the "capacity of an organization to purposefully create, extend or modify its resource base"(Helfat, 2007). Teece, Pisano, and Shuen (1997) define dynamic capabilities as "the firm's ability to integrate, build and reconfigure internal and external resources to address rapidly changing environments, indicating the importance of within firm dynamics for

growth and survival". These dynamic capabilities enable firms to improve their resources, learn to produce a given product, and/or diversify. Diversification may improve firm success prospects by reducing risk and keeping alive options in one market should other markets decline. Studies have shown that the ability to change the business model increases firm survival (Cucculelli and Peruzzi, 2020). Initial conditions may matter here as well, as firms that start with more "learning capacity" (reflected in the human capital of the founders and employees) are better able to learn and more likely to be successful. Better managerial abilities translate into lower costs at any given size, and these lower costs lead firms to choose to operate at a large scale (Lucas Jr, 1978). Better management practices of firms translate into higher productivity levels (Bloom, Sadun, and Van Reenen, 2016) at any given size, and these higher productivity levels may lead firms to choose to operate at a large scale. For new firms, this means "scaling up" before productivity levels increase, given that it takes some time for new firms to recruit and absorb employees (Penrose, 1959; Garnsey, 1998). Besides the direct effect of human capital (via managerial abilities) upon productivity and, thus, survival, a high stock of human capital in the firm may also be a consequence of some other intrinsic firm ability, as high-quality firms find it easier to attract highly skilled employees. Earlier studies have found human capital to be a good predictor of survival (José Mata and Portugal, 2002; Gimeno et al., 1997; Cooper, Gimeno-Gascon, and Woo, 1994; Unger et al., 2011). The observed size of firms can thus result from their superior managerial ability. Empirical tests of this capability learning theory have shown that (new) firms that invest in new knowledge are more likely to survive and grow (Stam and Wennberg, 2009; Ugur and Vivarelli, 2021).

These two theories of firm dynamics can be tested with specific hypotheses. Our null hypothesis is that the success of young firms is random. Based on these two theories, we formulate the following hypotheses:

- Hypothesis 1 (noisy learning): The success of young firms depends on the early-life characteristics of the firm.

- Hypothesis 2 (capability learning): The success of young firms depends on the changing firm characteristics during the life course.

We formalize these two theories in Section 2.3.

2.3 Methodology

We conceptualize firm success at a given age as a two-way process involving survival and size attainment. Survival is a necessary but not sufficient condition for success. We define the economic success of firm i at age a as reaching a specific turnover size in real terms, conditional on having survived up to that point. The threshold is determined using a specific quantile from the turnover distribution, which is then applied to create the economic success indicator as follows

$$S_{i,a} = \begin{cases} 1 & \text{if turnover} \geq q_\alpha \ \& \ I_{i,a-1} = 1 \\ 0 & \text{otherwise} \end{cases}$$

where the survival indicator $I_{i,a-1}$ equals 1 if the firm has survived for $a-1$ years and q_α is the quantile of order α from its distribution in 2019. We use all firms in 2019 as the benchmark year to calculate the cutoff, as it is less likely to be impacted by business cycle fluctuations, such as those following the 2008 crisis or the disruptions caused by the 2020 COVID-19 pandemic. This choice ensures a stable and accurate reference point for our analysis. In our analysis, we set $\alpha = 0.2$, corresponding to a turnover threshold of approximately 150 thousand euros, to define whether a firm makes a minimally meaningful contribution through its operations. This cutoff captures the lower bound of successful performance while maintaining a balanced approach and avoiding extremes that could bias the analysis. For robustness, we run sensitivity checks using a broader range of quantiles and applying sector-specific cutoffs $q_{\alpha,s}$, where s represents the industry of firm i.

2.3.1 Conceptual framework

Having defined our measure of economic success, we apply the framework developed in Mueller and Spinnewijn, 2023 for the status of unemployment to the status of the economic success of firms. Let $T_{i,a}$ represent the (latent) economic success type of startup i at age a, defined as the (unobserved) probability that its turnover exceeds a threshold q that is a function of the founding conditions X_i and an idiosyncratic error $\varepsilon_{i,a}$.

$$T_{i,a} = T_a(X_{i,1}) + \varepsilon_{i,a} \tag{2.1}$$

Assume that firm success, denoted by $S_{i,a}$, is a Bernoulli random variable that depends on the success type $T_{i,a}$, which defines whether firm i will be successful at age a, conditional on having survived up to age $a-1$. It is defined as follows

$$S_{i,a} = \begin{cases} 1 & \text{with probability } T_{i,a} \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

In expectation, any difference in success probability types for firms aged a and $a' < a$ generates an equivalent difference in observed survival rates: $E_a[S_{i,a}] - E_{a'}[S_{i,a'}] = E_a[T_{i,a}] - E_{a'}[T_{i,a'}]$. The expectation subscripts define the set of firms that are used to calculate the expectations. That is, $E_a[T_{i,a} - T_{1,a'}]$ represents the expected difference in the probability of success between ages a and a' computed for firms that have survived up to age a. This difference in success probability types can be decomposed into a part driven by dynamics within the firm and a part driven by dynamic selection

$$E_a[T_{i,a}] - E_{a'}[T_{i,a'}] = \underbrace{E_a[T_{i,a} - T_{i,a'}]}_{\text{within-firm dynamics}} + \underbrace{E_a[T_{i,a'}] - E_{a'}[T_{i,a'}]}_{\text{dynamic selection}} \tag{2.3}$$

$$= E_a[T_{i,a} - T_{i,a'}] + \frac{cov_a(T_{i,a}, T_{i,a'})}{1 - E_a(T_{i,a})} \tag{2.4}$$

The first term in equation 2.3 compares the average success probability across two different ages, a and a', for the group of firms that survive to

age a. The second term compares the average probability of success at a given age for two different groups of firms: those that have survived until a and those that have survived until a'. Equation 2.4 shows that the dynamic selection effect is driven by the heterogeneity in the success probability types of firms that is persistent, as measured by the covariance in the probability of success at ages a and a'. Obviously, individual heterogeneity in success probability types is a necessary condition for dynamic selection to impact the age-differential success rates. However, it is not sufficient, as individual heterogeneity needs to be persistent for dynamic selection to take place: without persistency, success probabilities at age a' would be orthogonal to success probabilities in year a, which rules out dynamic selection of 'good firms.'

Note that all the terms in both equations 2.3 and 2.4 are latent, but we observe whether a startup is successful at a given age. We propose a prediction model $f_a(\cdot)$ that uses founding conditions to explain success

$$S_{i,a} = f_a(X_{i,1}) + e_{i,a}, \tag{2.5}$$

where the error $e_{i,a}$ originates from random shocks to success probabilities (equation 2.1) and sampling noise (equation 2.2). Mueller and Spinnewijn, 2023 prove that under unbiasedness, i.e., $E_a[\hat{S}_{i,a'}|X_i] = T_{i,a}$, it holds that:

$$cov_a(S_{i,a}, \hat{S}_{i,a'}) = cov_a(T_{i,a}, T_{i,a'})$$

This allows us to estimate and infer persistent heterogeneity in the model, as well as, to identify an upper bound for the within-firm dynamics in success probability types between ages a and a'. They also provide a lower bound for dynamic selection term in equation 2.4 as

$$\frac{cov_a(T_{i,a}, T_{i,a'})}{1 - E_a(T_{i,a})} \geq \frac{cov_a(S_{i,a}, \hat{S}_{i,a'})}{1 - E_a(S_{i,a})} \tag{2.6}$$

Mueller and Spinnewijn, 2023 show that this inequality is binding when unobserved heterogeneity is purely transitory or when all of the firm heterogeneity in success is predicted by the model.

2.3.2 Estimation strategy

We estimate the relationship between a firm's early-life conditions and its economic success in equation 2.5 by analyzing firms that entered the market between 2006 and 2020 in the Netherlands. To assess the initial conditions of firm i (X_i), we examine a high-dimensional set of startup characteristics. Since the functional form of $f_a(\cdot)$ is unknown, we use extreme gradient boosting (XGBoost) to estimate it. XGBoost is a tree-based ensemble method that operates within a gradient boosting framework. It has been widely recognized for its exceptional competitiveness and reliability in various prediction tasks, primarily due to its efficient use of computational time and memory resources (Gumus and Kiran, 2017; Abbasi et al., 2019; M. Li, Fu, and D. Li, 2020). It is particularly well-suited for this application due to its ability to handle large datasets with high dimensionality and its robustness in managing unbalanced classes, which are common challenges in predicting startup survival and success. Additionally, its robustness to missing values allows it to effectively handle various patterns of missing data, especially in a firm's financial accounts (Falco J Bargagli-Stoffi, Incerti, et al., 2024a).

The characteristics in $X_{i,1}$ that we consider include firm financial information (from balance sheets and profit and loss statements), firm-specific indicators (such as sector, size, and age), workforce characteristics (including age, gender, migration background, and educational level), and owner characteristics (age, gender, migration background, educational level, and wealth). Additionally, we include control variables like the exit rate at the industry-region level and adjust all euro-based variables for inflation. The complete list of variables used in the analysis is provided in Table B.4.1 in the Appendix.

We split the dataset of 283,485 firms into three mutually exclusive subsets: a training set including 20% of the observations (56,697 firms), a validation set with 10% (28,348 firms), and a testing set containing the remaining 70% (198,441 firms). Each firm, along with all its time-series, is uniquely assigned to one of these three sets. In other words, the random assignment is done at the firm level—not at the level of individual

observations. This ensures that no single firm contributes data to more than one subset, thereby avoiding temporal and structural inconsistencies.

We train XGBoost to estimate the probability of success, using the training set. Next, we use the trained XGBoost model to predict the success probabilities for startups in the validation set. These predicted probabilities are then used to estimate a linear spline with ten segments, using the same validation set. The linear spline estimated from the validation set, combined with the XGBoost model trained on the training set, is then used to produce the final predictions for the test set. We use a wide test set to ensure robust out-of-sample analyses and to obtain reliable insights into prediction performance. Thus, the test set is reserved for assessing the model's out-of-sample predictive accuracy and for exploring the heterogeneity in predicted success probabilities among out-of-sample startups. Specifically, each test observation is first passed through the XGBoost model to compute a success probability, which is then fed into the spline model to generate the final prediction.

To predict future economic success using early-life features, we estimate various models across different ages. While the startups allocated to the training, validation, and test sets remain consistent across models, the actual startups included at each age vary. This variation arises due to the natural attrition in the dataset and the inherent challenge of censoring affecting survival. As startups age, the sample size reduces because some startups exit the dataset. Moreover, survival is not always observed, not because the startup has ceased to exist, but because we lack the data to confirm its current status. This combination of attrition and censoring means that as we look at older startups, the sample not only becomes smaller but also more selective, often including only those that have either survived or for which survival status is known. This reduction in sample size underscores the importance of carefully interpreting results from models focused on later ages, as they reflect a progressively narrower subset of the original population.

We employ three key metrics to evaluate our models: the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), the Area

Under the Precision-Recall Curve (AUC-PR), and the R^2. The AUC-ROC evaluates the model's ability to distinguish between economically successful firms and those that are not, providing a robust measure of classification accuracy across varying thresholds. The AUC-PR focuses on the model's ability to identify the positive class (successful firms), emphasizing two critical components: precision (the proportion of predicted successes that are correct) and recall (the proportion of actual successes that are identified). In scenarios with strong class imbalance, AUC-PR offers a more reliable assessment by mitigating the misleading influence of the large negative class. The R^2 measures the proportion of variance in the outcome explained by the model, highlighting its overall fit and the predictive strength of early-life features. Additionally, R^2 provides insights into the lower bound of dynamic selection, as we will detail later.

As machine learning models become more sophisticated and opaque –often referred to as "black boxes"-— it is crucial to ensure that these models are not only accurate but also interpretable and transparent. We therefore use Shapley values (Rozemberczki et al., 2022) and ALE plots (Apley and J. Zhu, 2020) to interpret our models, see appendix B.1 for a detailed description of these methods.

2.3.3 Empirical tests supporting the theories of firm learning

We use the predictions of our models to quantify the heterogeneity and dynamics in firm economic success probability types and interpret them using noisy learning and capability learning as follows.

First, we consider the *heterogeneity in firm success types*. In particular, we use the ratio of the R^2 at age a over that of age a' to inspect the persistency of early life conditions in explaining observable heterogeneity in firm success types as firms age. Mueller and Spinnewijn, 2023 proof that this ratio measures the persistence of observable heterogeneity

$$\frac{R^2_a(\hat{S}_{i,a'}, S_{i,a})}{R^2_a(\hat{S}_{i,a}, S_{i,a})} = \frac{Var_a(T_{i,a'})}{Var_a(T_{i,a})} \tag{2.7}$$

where $R^2_a(\hat{S}_{i,a}, S_{i,a}) = \frac{COV_a(\hat{S}_{i,a}, S_{i,a})^2}{Var_a(\hat{S}_{i,a}) Var_a(S_{i,a})}$.
We then apply the decomposition in equation 2.3 to interpret the relative importance of noisy learning in explaining the increase in success as firms age. In particular, when the change in success rates between ages is, by and large, driven by dynamic selection, firm learning is more likely to be driven by noisy learning, although capability learning does not rule out moderate levels of dynamic selection. If, on the other hand, within-firm dynamics are, by and large, the most important driver, noisy learning is unlikely to be an important way by which young firms learn.

Second, we consider *the dynamics in success probabilities over the life-course of firms*. We start by imposing structure on our estimates of firm success and estimate the observable heterogeneity in firm success dynamics. In particular, we impose a log-linear relationship between firm success probabilities at age a– predicted using the early life conditions of the firm– and their age as follows

$$log(\hat{s}_{i,a}(X_{i,1})) = \beta_{i,0} + \beta_{i,1} a + \nu_{i,a} \tag{2.8}$$

Note that the imposed log-linear relationship allows interpreting the parameters as a Cox proportional hazards model in which $\beta_{i,0}$ equals the log of the baseline hazard function and $\beta_{i,1}$ is the proportional effect of age on the probability to be economically successful. To account for survival bias, this equation is estimated regardless of firm survival or economic success. We do this for a random sample of five thousand startups in the test sample.

As the dependent variable in the specification is an estimate, the distribution of $\beta_{i,0}$ and $\beta_{i,1}$ reflects both true dispersion as well as sampling error. Following Mueller and Spinnewijn, 2023, we shrink the coefficients for $\beta_{i,0}$ and $\beta_{i,1}$ using the sample means and sample variances as follows

$$\tilde{\beta}_{g,i} = (1 - w_i) E[\hat{\beta}_{g,i}] + w_i \hat{\beta}_{g,i} \quad \text{with} \quad w_i = \sqrt{\frac{VAR[\hat{\beta}_{g,i}] - \hat{se}^2_{g,i}}{E[\hat{\beta}_{g,i}]}}, g = 0, 1$$

Here $E[\hat{\beta}_{g,i}(X_1)]$ and $VAR[\hat{\beta}_{g,i}(X_1)]$ are the sample mean and sample variance of the estimated parameters before shrinkage. $\hat{\beta}_{g,i}(X_1)$ and

$\hat{se}_{g,i}(X_1)$ are the estimated coefficients and standard errors. The formula essentially shrinks the estimated coefficient for each firm to the sample average, where the shrinkage weight w_i increases with the standard error of the estimated parameters. As w_i is undetermined when $\hat{se}_{g,i}(X_1) > VAR[\hat{\beta}_{g,i}(X_1)]$, we set it equal to zero in this case.

We then use the estimates $\tilde{\beta}_{g,i}$ to empirically test for a relationship between the dynamics in economic success types and early life conditions. In particular, we estimate the following regression using LASSO (where the model is trained using 10-fold cross-validation)

$$\hat{\tilde{\beta}}_{i,1} = g(X_{i,1}) + v_i \tag{2.9}$$

This analysis allows us to test the importance of capability learning in explaining the rise in firm success over the life course. If we observe that firm-level characteristics on the financial position, the quality of the workforce, or the quality of firm owners are important in explaining firm success dynamics, we interpret this as evidence that firm-specific capabilities explain the success of firms. If, however, regional and time indicators are most important, this lends credibility to the belief that firms' economic success is explained by noisy learning. Naturally, this approach relies on the possibility of detecting a firm's capability from the measured firm characteristics. We observe high-dimensional data on the financial position of firms and important workforce and owner characteristics, like the wage of the workforce and age and wealth of the owners, as well as indicators for the regional economic environment in which the firm operates. Although it is likely that we do not observe all capabilities that contribute to the success of firms, we feel confident that we observe sufficient and detailed information at the firm level to pick up the correlation with these unobserved capabilities. The fact that we are able to explain economic success very well also at high ages indicates that we observe important features for economic success and supports our claim.

Third, we consider the *heterogeneity in firm success types over the business cycle*, which allows us to quantify to what extent annual variations in the success rate of young firms are driven by changes in their com-

position. We also consider whether the relationship between firm early life conditions and economic success is persistent. This is informative on whether capability learning or noisy learning is the dominant process by which learning takes place because capability learning stresses the importance of firm capabilities to respond to changing business environments. Noisy learning, in contrast, views firms as passive units in a changing economic environment.[1] Therefore, if we find evidence that the relationship between firm early life conditions and economic success change over time, and more so when time periods are further apart, this can be interpreted as evidence that firms adapt to changes in the environment, which would be in line with capability learning. If, however, the relationship between firm composition and firm success is stable across years, changes in the business environment do not call for a differential response by firms, which suggests that noisy learning is a dominant learning strategy.

We propose the following method to inspect whether the relationship between early life conditions and firm economic success is stable over time. To explain the method, we introduce additional notation for predictions of the success probability types

$$\hat{S}_2^m(X_1^{t-1})$$

here the superscript m denotes the modelled year, specifically the year in which the outcomes used to train the model were observed, and the subscript 2 indicates that success probability types are estimated for firms that are two years old. The superscript $t-1$ denotes the year when early-life conditions, X_1, were observed. The subscript 1 indicates that the firm was one year old at that time. From now on, we drop the subscript that denotes the firm's age and the age of early life conditions for readability. In this framework, we first set $m = t = 2015$ to estimate the probability of success at age 2 in 2015 using firm early-life conditions from 2014. This is formally expressed as $\hat{S}^{2015}(X^{2014})$. This dataset includes approximately

[1]Note that our sample periods contain the great financial crisis and the following recession, as well as the first year of the COVID-19 pandemic, and therefore the economic environment was not stable in the period we consider.

10,000 startups in the combined training and validation set.[2]

Next, we compute $\hat{S}^{2015}(X^{t-1})$ with $t = 2007, ..., 2020$. By keeping the prediction model $\hat{S}_2^{2015}(\cdot)$ fixed, we can examine how the (counterfactual) expected probability of success varies for startups that are two years old in year t. This probability is given by

$$E[\hat{S}^{2015}(X^{t-1}) \mid t = k], \quad k \in \{2007, \ldots, 2020\} \tag{2.10}$$

and captures changes in the composition of the startup pool over time. By comparing this counterfactual success rate to the actual success rate, we can inspect how much variation in the observed success rate of young firms can be explained by variation in the observable composition in early life conditions, evaluated using the 2015 economic model.

We use a related approach to measure the extent to which early life conditions that successfully explain economic success in one year can also explain it in other years. Once the model $\hat{S}^m(\cdot)$ is trained for a given year m, we use it to generate out-of-sample predictions for the test set. Specifically, we apply $\hat{S}^m(X^{t-1})$ for years $t = 2007, \ldots, 2020$. We then compute the R^2 between the observed success rate of firms that are 2 years old in year t and their (counterfactual) probability of success, using the model trained in year m. This is given by:

$$R^{2,t,m}(S^t, \hat{S}^m(X^{t-1})) = \frac{COV\left(S^t, \hat{S}^m(X^{t-1})\right)^2}{Var\left(S^t\right) Var\left(\hat{S}^m(X^{t-1})\right)}, \tag{2.11}$$

We compute equation 2.11 for all combinations of m and f and we interpret it as follows. When $m = t$, the R^2 represents the share of variation in observed success probabilities types for firms that are two years old in year t that can be explained by the relationship between early-life conditions and success in the same year. In contrast, when $m \neq t$, the R^2 measures the share that can be explained using the relationship between early-life conditions and success probabilities from a different year. In

[2]To increase the number of observations used for training the model, we use 50 percent of the observations in each year for training and 20 percent for debiasing.

that respect, we compute the ratio between the R^2 of cohort m' and that of our benchmark cohort $m = t$ as

$$\frac{R^{2,t,m'}(S^t, \hat{S}^{m'}(X^{t-1}))}{R^{2,t,m}(S^t, \hat{S}^{m}(X^{t-1}))}, \quad m = t \tag{2.12}$$

Note that the predictions of these R^2 are based on the same set of early life conditions. The only difference between them is the year used to establish a relationship between early-life conditions and firm success outcomes. Therefore, values close to one suggest that the relationship between early life conditions and firm success in estimation years m and m' are very similar. Conversely, values far from one suggest that the relationships between early life conditions and firm success have differed substantially over the years. This directly allows us to quantify to what extent relationships between early life conditions and firm success are stable over the business cycle.

2.4 Data

We identify firms from an annual unbalanced panel dataset at the firm level that includes the universe of all incorporated (limited liability firms and public liability companies) firms registered in the Netherlands from 2006 to 2021. We thus exclude sole proprietors, as these are more likely to reflect self-employment instead of (young) firms. This data was compiled by merging more than one hundred administrative datasets provided by Statistics Netherlands. A unique company identifier was used to track each firm over time using the General Company Registry (ABR), in which we define the creation and exit of firms based on their registration and deregistration data. We define start-ups as companies that are one year old and have submitted their annual financial statements. This final dataset contains specific details on the sector, age, and type of exit of the companies, whereby we distinguish between an exit due to a cessation of activity (shutdown), an exit due to a merger and acquisition, and an exit due to a restructuring of the firm. Subsequently, this dataset is integrated with information on firm balance sheets and corporate profit

and loss accounts (NFO), which provide detailed financial information ranging from turnover and balance sheet size to results from equity investments and trade credits. Finally, we merge the data sets of all taxable employment contracts (SPolis). In this way, we can bring together characteristics that provide information about the employees and owners of companies, such as their (average) age, gender, migration background, education level, and – for owners – their financial wealth. Figure 11

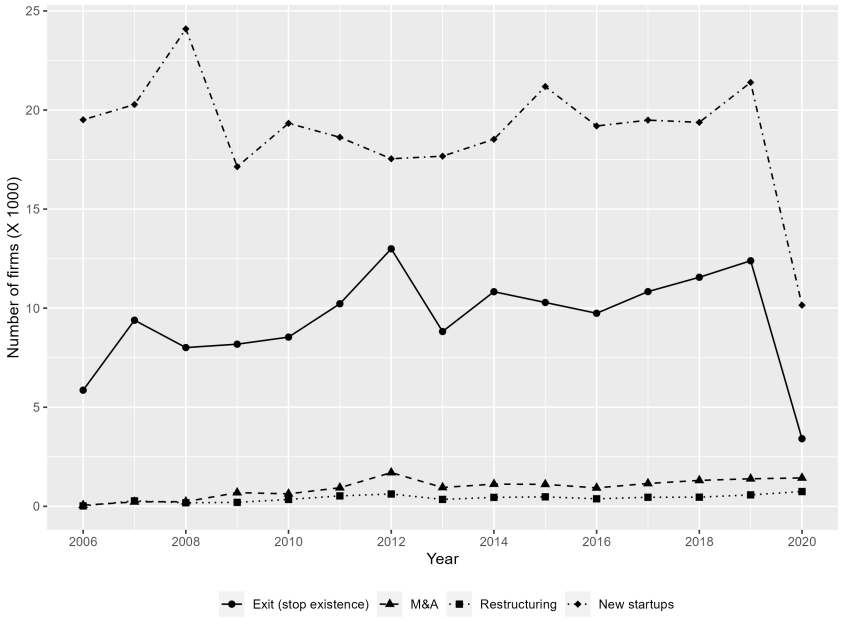


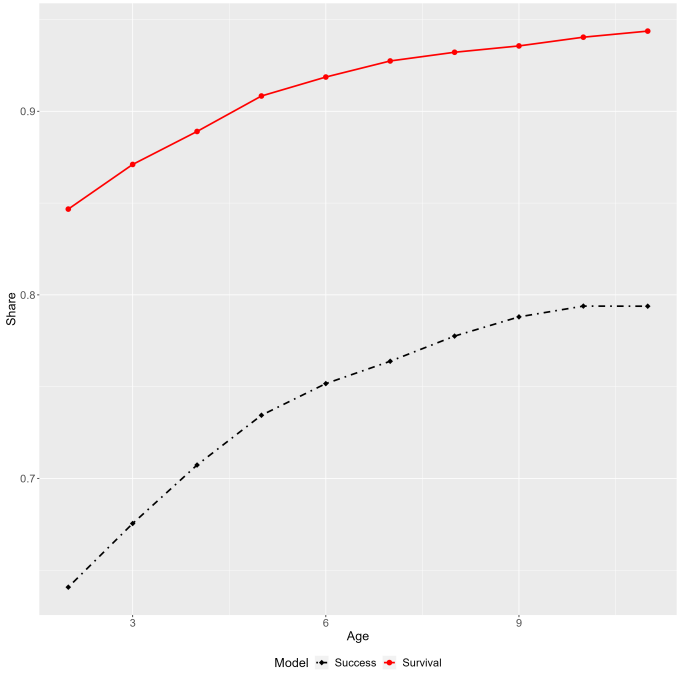
Figure 11: Exit vs. entry dynamics

shows that the number of firms in our dataset increases when the number of start-ups exceeds the number of exits, with exits falling into three categories: cessations (Stops), mergers and acquisitions (M&A), and restructurings. The number of stops is subject to considerable fluctuation and rose sharply between 2008 and 2012, probably due to the effects of the financial crisis. After a period of relative stability, the number of stops fell sharply in 2020, probably due to the pandemic and the associated

government interventions. In contrast, M&A and restructuring remain relatively stable over time, with both categories showing a gradual upward trend. This pattern could indicate increasing market consolidation as firms opt for mergers or restructurings rather than a complete exit. Meanwhile, the number of new firms declines after the financial crisis, but stabilizes in the following years. Similar to stops, there was also a sharp decline in start-ups in 2020, which is likely due to the disruption caused by the COVID-19 pandemic. Moreover, it shows that the number of established companies is growing faster than the number of start-ups, which leads to a decline in the entry rate, especially towards 2020 (see table B.2.1 in the Appendix). The number of start-ups peaked in 2008 before falling after the crisis. It then fluctuated with another peak in 2015, followed by a sharp decline in 2020, probably due to the pandemic. On the other hand, the number of established companies is growing steadily, indicating an expanding market. Figure 12 depicts the proportion of the 198,441 startups in the test set that survive and the proportion of those that achieve economic success — defined as survival and reaching a turnover threshold of €150,000 — as a function of their age. The figure shows the well-documented positive relationship between firm age and survival. At the age of two years, for example, around 85 percent of companies survive until the next year. This rate rises to 92 percent at the age of six years and reaches 94 percent when the companies are ten years old. While the relationship between age and economic success is also positive and monotonic, the economic success curve is consistently below the survival curve, suggesting that economically successful firms are a subset of surviving firms. At the age of two, 64 percent of companies are successful; by the age of six, it is already 75 percent.

The relationship between economic success and age changes with the selected cut-off value. If the cutoff is reduced, the line shifts upwards and approaches the survival rate (the two lines are identical when the turnover cut-off value is zero). If the cut-off value is increased, the line shifts downwards. Figure B.5.2a in the Appendix shows the survival rate and the economic success rate in the test set, plotted against age. The cutoffs used correspond to the fifth, tenth, twentieth, and fortieth

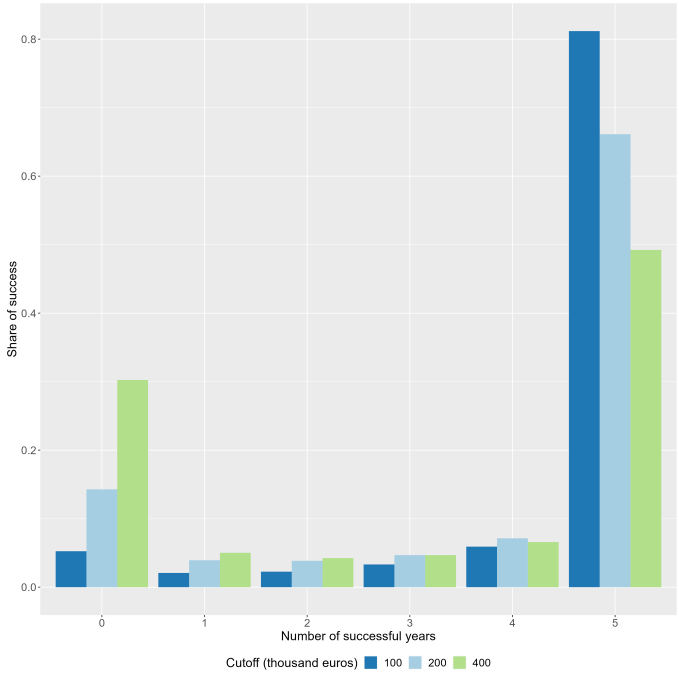
Figure 12: Survival and economic success



percentiles of turnover in 2019. All cutoffs yield a similar relationship between economic success and age. This pattern for economic success is consistent across cohorts (Figure B.5.2b in the Appendix). In addition, Figure B.5.3 shows that the relationship between economic success and age varies by sector, suggesting that in some sectors, more firms reach a turnover threshold of €150k at a given age.

The economic success of young firms varies greatly. Most firms are either always successful or never successful in the first five years of their existence. This distribution is affected by the threshold value chosen to determine economic success, as shown in Figure 13. The figure shows that over 80% of the startups in our sample that survive the first five years also achieve a minimum annual turnover of €100,000 in each of these years. In contrast, only 5% do not manage to be successful within

Figure 13: Success rate by years of success and cutoff



Notes: The plot shows the share of successful startups among those that survived the first five years. A total of 123,155 startups survive over this period.

the first five years. A marginal group of startups alternates between success and unsuccess conditions, achieving success in only 1 to 4 years. When applying higher turnover thresholds of €200,000 and €400,000 per year, only 66% and 49% of startups achieve success each year. Table 7 lists the most important economic indicators for startups at the beginning of their life course and five years later for those that have survived. This comparison illustrates the significant changes that occur as startups mature and face the challenges of early business development. In their first year of existence, startups generally operate on a modest scale, with a median total balance sheet value of 291,000 euros. By the fifth year, this

median figure almost doubles, indicating that these companies are expanding and growing their business. There is also a significant increase in tangible assets, reflecting the accumulation of physical resources as startups expand their operations. This growth suggests that startups are likely to invest more in tangible assets as they stabilize and expand. Successful startups not only survive but thrive by scaling up their operations and assets. Intangible assets remain at a similar level in all statistics, suggesting that the relative importance of non-physical assets such as patents or software remains the same over time. The analysis of debt and equity dynamics reveals that startups generally improve their financial stability as they mature. The mean equity increases from 390 in the first year to 630 thousand euros in the fifth year, indicating a stronger capital base. At the same time, the leverage ratio falls from 0.68 to 0.56, and the debt-to-equity ratio from 0.75 to 0.64, indicating that firms are either reducing debt or increasing their equity faster than their liabilities. These changes indicate improved solvency and better financial management over time, as shown by the increase in the mean solvency ratio from 0.59 to 0.8.

The operating performance, measured in terms of sales and EBITDA, also shows significant improvements. The median turnover increases from 373 thousand euros in the first year to 628 thousand euros in the fifth year. This growth in turnover indicates that the companies are expanding their market presence and customer base. EBITDA, a key measure of operational profitability, moves from a slightly negative mean in the first year to a positive figure of 25 thousand euros in the fifth year, highlighting the shift towards profitability as the companies mature.

Overall, the statistics illustrate the significant transformation that startups undergo in their early years. Initially, they are characterized by high variability in terms of size, financial health, and operational performance. However, those that survive to the fifth year tend to have a stronger financial structure, better operational efficiency, and greater market presence. In the Appendix, Tables B.2.2 and B.2.3 provide further statistics about startups' workforce-based features.

Table 7: Startups characteristics

(a) Startups in their first year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Total balance sheet	1109.23	290	2803.83	6.1	22194.91	Thousand €
Tangible assets	243.1	10.88	876.14	0	7195.34	Thousand €
Intangible assets	18.2	0	73.81	0	556.48	Thousand €
Liquid assets	161.74	31.58	454.63	-0.12	3746.85	Thousand €
Equity	386.66	63.2	1390.64	-1449.06	11186.62	Thousand €
Long term debt	249.56	0	858.26	0	6801.2	Thousand €
Short term debt	253.31	58	702.91	0	5734.12	Thousand €
Turnover	2022.17	371.41	6130.74	7.9	51628.43	Thousand €
Result before taxes	84.99	17.71	348.66	-757.34	2647.09	Thousand €
Taxes	-16.62	-1.21	50.92	-417.91	18.33	Thousand €
Value added	363.78	122.69	874.68	-85.87	7182.83	Thousand €
EBITDA	-0.14	2.35	270.06	-1399.13	1502.94	Thousand €
Dividends	21.12	0	113.06	0	977.83	Thousand €
Labor productivity	58.9	31.35	110.11	-30.74	781.15	Thousand €
ROA	0.02	0.06	0.44	-1.98	0.94	Ratio
Solvency ratio	0.6	0.08	2.56	-7.52	16.28	Ratio
Debt ratio	9.46	2.21	27.54	0	216.62	Ratio
Age	1.52	1.5	0.28	1.08	2	Number of years
Entry year	2012.76	2013	4.26	2006	2020	Year
Year	2013.76	2014	4.26	2007	2021	Year

(b) Startups in their fifth year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Total balance sheet	1509.22	531	3133.54	8	23692.05	Thousand €
Tangible assets	316.88	24.29	965.78	0	7566.16	Thousand €
Intangible assets	18.29	0	75.5	0	575.19	Thousand €
Liquid assets	242.51	67	539.7	-0.08	4122.46	Thousand €
Equity	629.66	193.73	1605.04	-1412.45	12125.38	Thousand €
Long term debt	288.38	0	892.32	0	6827.78	Thousand €
Short term debt	326.89	78.7	817.14	0	6293.29	Thousand €
Turnover	3048.78	628	7785.23	9.08	60578.68	Thousand €
Result before taxes	140.55	37.94	424.89	-701.08	3058.34	Thousand €
Taxes	-24.52	-3.64	65.08	-498.82	0.3	Thousand €
Value added	561.81	209.47	1125.46	-77.82	8470.09	Thousand €
EBITDA	25.34	6.32	308.21	-1326.51	1734.45	Thousand €
Dividends	35.98	0	145.42	0	1162.25	Thousand €
Labor productivity	61.34	40.7	81.94	-28.01	581.02	Thousand €
ROA	0.05	0.07	0.33	-1.85	0.92	Ratio
Solvency ratio	0.8	0.12	2.84	-7.34	17.36	Ratio
Debt ratio	11.85	3.04	29.28	0.01	221.89	Ratio
Age	5.53	5.5	0.28	5.08	6	Number of years
Entry year	2010.99	2011	3.18	2006	2016	Year
Year	2015.99	2016	3.18	2011	2021	Year

Notes: summary statistics for 283,486 firms during their first year of classification as startups. Additionally, Table 7b provides summary statistics for the subset of startups that successfully survived for 5 consecutive years, comprising 123,155 out of the initial 283,486.

2.5 Results

2.5.1 Performance and calibration

Overall, our models demonstrate strong predictive performance for economic success, regardless of the age of the firm or whether we measure economic success using the pooled or sector-specific cutoffs (see Table 8). The R^2 for economic success at age two exceeds fifty percent, drops as firms get older, but is still 23 percent when firms are ten years old. Both AUC-ROC and AUC-PR exceed 0.90 at age two, indicating a high level of accuracy. Over time, AUC-ROC and R^2 gradually decline, reaching 0.79 and 0.24, respectively, by age ten, while AUC-PR remains relatively stable. This suggests that the model is particularly effective at identifying successful startups but less so at distinguishing between successful and unsuccessful firms. This is intuitive, as the model predicts that a large share of firms will reach the minimum revenue threshold of €150,000. However, as shown in Appendix Table B.4.4, when the success threshold is raised to higher quantiles, AUC-PR drops significantly below AUC-ROC. This occurs because increasing the cutoff makes the dataset more imbalanced, and it becomes harder to predict firms that are very likely to be successful.

The sector-specific model shows a nearly identical pattern to the pooled model, indicating that predictive power for economic success is relatively unaffected by differences in industry sizes or industry-specific circumstances. This initial robustness suggests that industry heterogeneity has a limited impact on the model's ability to predict economic success. When compared to the simple survival prediction (see Table B.4.2 in the Appendix), the superior performance of the economic success models can be attributed to the fact that economic success is more directly linked to observable early-life characteristics, enabling more accurate predictions. Apparently, the decision to withdraw a firm from the market is influenced by a broader range of factors, including unpredictable external shocks, than having economic success. Additionally, the declining performance across all models with increasing age underscores the chal-

lenges of long-term prediction as uncertainties accumulate and sample sizes shrink due to attrition and censoring.

Table 8: Performance metrics: economic success prediction models

(a) Pooled cutoff ($\alpha \approx 0.2$, $q_\alpha = 150,000$)

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9017	.9259	.5257	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.8120	.9079	.2906	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.7906	.9188	.2356	10	1

(b) Sector-specific cutoffs ($\alpha \approx 0.2$)

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.8974	.9246	.5147	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.8009	.9065	.2707	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.7623	.9124	.2053	10	1

We find that our machine learning method predicts economic success better than conventional estimators like logistic regression or OLS. In Appendix B, we provide evidence that this is due to the flexibility of the model, particularly the ability to accommodate non-linear effects of features. Moreover, Tables B.4.4 and B.4.6 present the performance results for predicting economic success across a wider range of cutoffs, showing that the relatively good performance of our method is not specific to the cutoff chosen. From this point forward, unless stated otherwise, we will use the pooled cutoff model to estimate the probability of success.

In contrast to early life conditions, contemporaneous conditions – firms' characteristics from the previous year– remain strong predictors of success regardless of age (Table 9), pointing towards the key role played by up-to-date information in predicting economic success the next year. The R^2 increases from about fifty percent at age two to about sixty percent at age ten. One explanation for this pattern is that the sample becomes more strongly concentrated around high success types, which reduces sample noise. Notwithstanding the difference in predictive performance when models are given the early life conditions or contemporaneous features, we note that predictions by both models are similar. The correlation between predictions of economic success at age six when features are measured when firms are one or five years old is 0.78 percent. Also, a bivariate regression of these predictions, where predictions from

features when firms are five years old is the dependent variable, has a precisely estimated slope coefficient of 0.93.

When we predict economic success at ages six and ten using both contemporaneous and early life conditions, the predictive performance is virtually identical to that reported in Table 9.[3] As predictive performance does not change significantly, early life conditions do not carry *additional* information that is predictive of economic success over the information that is contained in contemporaneous features. This also suggests that our main models, which use only early life conditions to predict economic success, are able to predict economic success later in life mainly because initial life conditions are strongly correlated with firm characteristics later in life.

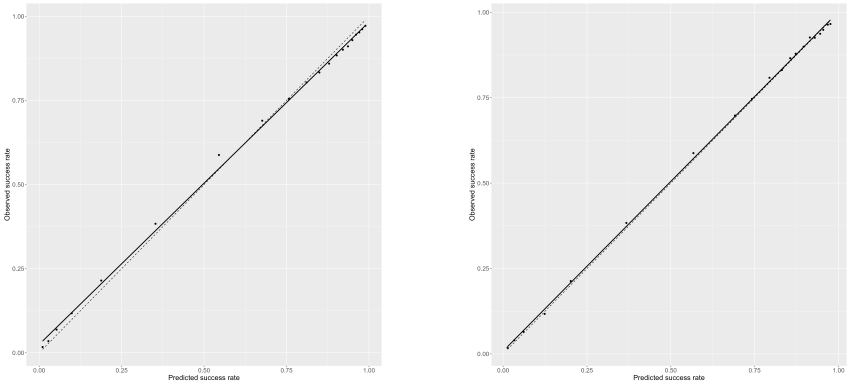
Table 9: Economic success models using contemporaneous features ($\alpha \approx 0.20$, $q_\alpha = 150,000$)

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9017	.9259	.5257	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9144	.9568	.5715	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9147	.9648	.5977	10	9

The finding that early life conditions do not affect firm economic success conditional on contemporaneous features of the firm contrasts sharply with the results of Geroski, José Mata, and Portugal (2010), who find that early life conditions have a significant, yet fading, impact on firm survival later in life conditional on contemporaneous features. As we will detail later, we do not find evidence that significant predictors in the analysis by Geroski, José Mata, and Portugal (2010) are important drivers of economic success when our high-dimensional set of firm financial characteristics is included in the feature space. This suggests that the inclusion of high-dimensional firm financial indicators in the feature space is the main reason why our results differ from Geroski, José Mata, and Portugal (2010).

[3] As they only differ from the third digit onwards, we do not report results here. They are available upon request.

Figure 14: Calibration plots for success probabilities at age 2



(a) First stage: XGB

(b) Overall model: 2S-XGB

Finally, the predictions by our machine learning model are well calibrated. Figure 14 shows the observed success rates against the predicted success rates of our model. The model predictions are grouped into 20 bins, each representing the mean observed and predicted success rates. The calibration plot for this combined approach shows a marked improvement, with the points adhering closely to the 45-degree line across all bins. Additionally, Figures B.5.4-B.5.6 in the Appendix show the calibration plots for various models and time horizons: these cover predictions of survival and success probabilities (using both pooled and sector-specific cutoffs), as well as survival predictions at different ages using features from the prior year.

2.5.2 Distribution of economic success probability types

Success types for startups at different ages differ substantially. We derive this conclusion from the density and complementary cumulative distribution function (CCDF) plots in Figure 15 computed for all startups observed in the test data when they are one year old.

The densities are bimodal at all ages, suggesting the existence of two distinct subpopulations of startups. Startups typically show higher ob-

servable success probabilities at a young age, indicating that those who succeed early in life tend to do so with a relatively strong likelihood. This could reflect that only the most promising startups—those with strong founding conditions, innovative products or competitive advantages—achieve early economic success. However, as firms age, the density shifts toward lower success probabilities. This change may be driven by increased competition, market saturation, or operational challenges that arise over time, making it more difficult to achieve or maintain the size threshold. Essentially, older firms experience a stabilization or decline of their growth potential, further reinforcing the idea of two distinct subgroups: one with high potential that succeeds early, and another that faces greater challenges as they age.

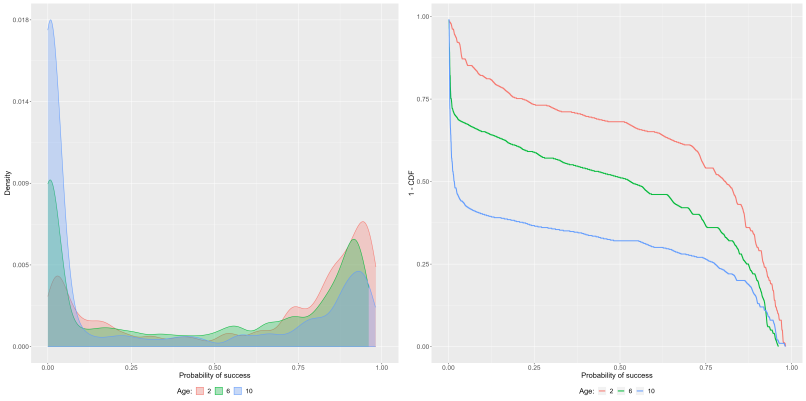
The CCDFs in Figure 15b support this by showing that, as firms age, fewer of them reach high success probabilities. This trend suggests that while firms that persist into later stages continue to operate, their chances of achieving economic success diminish over time.

Naturally, the share of firms with very low observed success probability types at age a increases as firms age because of sample attrition. However, we still predict survival probability types for these firms as we report predictions for all startups using their early life conditions. In Figures 15c and 15d, we report the density and CCDF plots of observed success probability types for firms aged two, six, and ten for long-run survivors, defined as startups that ex-post stay in the market for ten years. The density distribution in Figure 15c of this subset is very different from the one from the general population. At age two, the distribution of the success probability types for long-run survivors is very similar to the distribution for all firms. However, at later ages, the concentration at high observed success probability types has increased, whereas the concentration at lower values has diminished. Also, the distribution of success probability types at ages six and ten is very similar. As the sample remains constant, this suggests that some long-run survivors succeed in increasing their economic success probability types between ages two and six.

In the Appendix, we report estimated densities and CCDFs for both

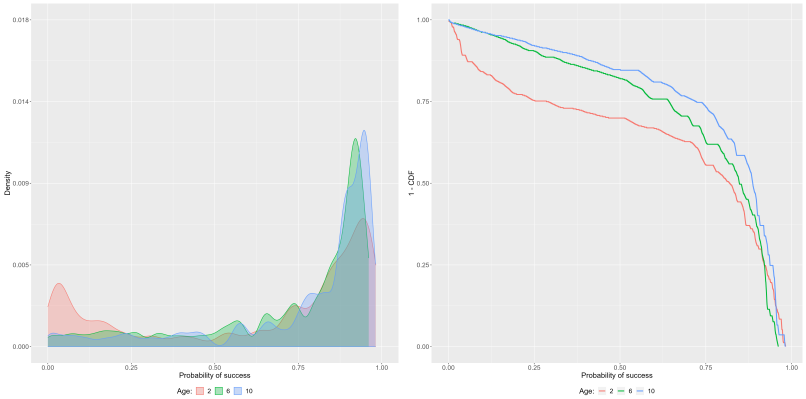
survival and sector-specific models. The sector-specific results closely align with those from the pooled model, while the survival models reveal distinctly different dynamics. Survival probabilities are left-skewed, with an increasing concentration at higher values as firms age. This trend is consistent with existing literature, which generally finds that survival probability increases almost 'monotonically' over a firm's lifespan. However, the dynamics of success and survival differ markedly and appear almost inverted. This suggests that the drivers of both short and long-term resilience, i.e., the ability to survive, are quite different from those that foster success and growth.

Figure 15: Success probabilities at different ages



(a) All testing startups (density)

(b) All testing startups (CCDF)



(c) Conditional on long-run survival (density)

(d) Conditional on long-run survival (CCDF)

2.5.3 Drivers of economic success

We describe how certain founding conditions contribute to the survival and economic success of startups at later stages. First, we do this by removing groups of characteristics from the model and assessing the impact on the model's explained variability. Next, we use Shapley values

to identify which variables contribute the most to the likelihood of economic success overall. Finally, we examine the marginal effect of revenue on the probability of survival and the likelihood of economic success in our models.

Group-wise decomposition of R^2

We inspect the importance of blocks of independent variables by considering the change in R^2 when these blocks are removed from the specification. This is informative on the importance of economic success because the R^2 drops when a group of variables is important, and their role in forming the predictions cannot be taken over by variables that are still included. We consider the following groups of variables and combinations of them: financial statement variables, worker and DGA variables and age of the firm, size of the firm (measured as the number of employees), industry sector in which the firm is active, and region in which the firm is located.

We find that financial statement variables are the most important group of variables. Without this group, the R^2 halves at almost all the ages of success are considered, see Table 10. In particular, the R^2 drops by 27.06 percentage points (from 52.57 percent to 25.51 percent) when this group of variables is removed. Next, the most important variables are the characteristics of the workforce and of the firm's owner. However, it's important to note that the R^2 drops by only 0.26 percentage points (from 52.57 percent to 52.31 percent), with a more noticeable drop at later ages. This suggests that these specific early-life characteristics have minimal predictive contribution early on, but their influence grows later (e.g., a 2.36 percentage-point reduction at age 10). The other features have minimal and often negative impacts on the explained variability of the model, suggesting these features may not strongly explain success in isolation or are redundant given other variables. Likewise, financial constraints have little to no predictive power, with reductions close to or below zero. Note that sometimes the R^2 *increases* when a group of variables is removed, indicating that the model better explains the target *out of sample* when this

group is removed.[4]

To sum up, the R^2 analysis underscores the dominant role of financial statement variables in predicting economic success, particularly at earlier stages. Conversely, workforce-related variables, while historically deemed important for survival, appear less critical when other factors are considered. The increasing importance of worker and DGA variables at later ages may indicate that workforce dynamics become relevant in explaining success further into a firm's life course. In the Appendix, we provide further analyses for the simple survival prediction and for the economic success model with sector-specific cutoffs. The models show that financial statement variables are by far the most important in all models. Conversely, despite the literature showing that workforce variables are important in explaining firm survival, our analysis shows that the predictive performance of these variables for survival is low, conditional on the other groups of features.

Table 10: Reduction in R^2 for economic success (pooled model, $\alpha = 0.2$)

Age to success	age 2	age 6	age 10
R^2 of the full model	52.57	29.06	23.56
percentage points reduction in out of sample R^2 when removing			
Financial statement variables and year	27.31	12.28	8.44
Financial statement variables	27.06	12.16	8.44
Year	0.13	-0.26	0.24
Worker and dga variables	0.26	0.94	2.36
Worker variables	0.25	0.03	0.53
Dga variables	0.10	0.11	-0.09
Age, size, sector, area	-0.09	-0.27	0.11
Age	-0.08	0.07	0.12
Size class (OECD)	0.10	-0.15	-0.28
Sector	0.10	0.17	0.27
Region	0.09	-0.20	0.07
Financial constraints variables	-0.09	-0.20	-0.13

[4]As is well known the in sample R^2 *always* increases when variables are added to the specification, but the out of sample R^2 might fall. Likewise, the in-sample R^2 always decreases when variables are removed, but the out-of-sample R^2 might increase, especially when machine learning and high-dimensional data are combined. When it happens, it points out that the model better predicts the target because noise is removed, making it easier for the model to find the relevant structure.

Most important predictors

We use Shapley Values to identify the most important predictors of startup success at different ages. For each firm in the test data, we calculate the Shapley values of all variables used to predict future success. Then, for each variable, we compute the average absolute Shapley value. The resulting Shapley plots share a uniform x-axis scale within the same model class (e.g., success prediction), facilitating straightforward comparisons while maintaining clarity in the visual representation.

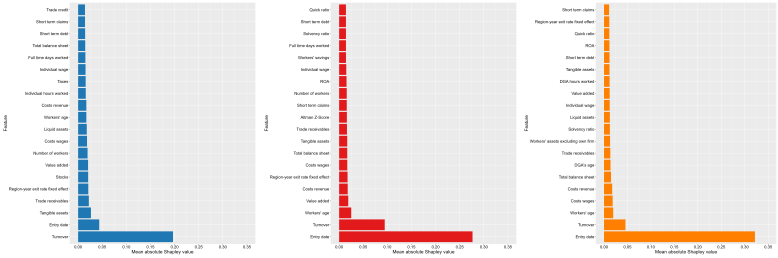
The analysis reveals that turnover and entry data are by far the most important in predicting the outcome, see Figure 16. This pattern holds both when the model's predictive power is lower at later stages and when it is higher at earlier stages. The importance of turnover suggests that turnover in the first year is a key predictor of whether a minimum turnover level will be achieved at a later stage. The relative importance of the entry date indicates that there are significant differences between companies that start during periods of economic boom or bust.

Figure 17 compares the Shapley values at age 2 for startups in the testing set with a high probability of success—those in the top quartile of the success probability distribution—with those with a low probability of success, falling into the bottom quartile.

The early-life conditions that most significantly contribute to explaining success are strikingly similar between the two groups. However, turnover plays a more prominent role in shaping the predictions for startups in the bottom quartile, which aligns with expectations. Startups in the top quartile are likely those capable of surpassing the success threshold imposed by the model. In contrast, startups in the bottom quartile face a greater challenge in achieving a viable size. They may need more effort to meet the size required for success, whereas top-quartile startups face a relatively lower marginal effort to keep their size.

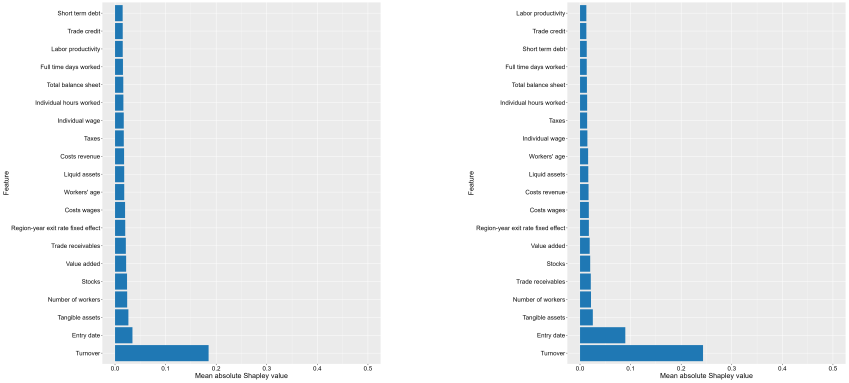
In the Appendix, we provide a corresponding comparison of early-life success predictions at ages 6 and 10. Once again, the timing of market entry emerges as the critical factor the algorithm exploits to distinguish between successful and unsuccessful startups.

Figure 16: SHAP comparison for success at different ages



(a) Age 2 **(b)** Age 6 **(c)** Age 10

Figure 17: SHAP comparison for success at age 2: top vs. bottom quartile firms



(a) Top 25% **(b)** Bottom 25%

Marginal effects of turnover

Given the importance of turnover in predicting economic success, we examine how predictions change with this variable. Figure 18 shows the effect of increasing turnover by €1000 on the probability of being economically successful. We consider how this effect changes as firms get older. The marginal effect is plotted as a step function in which the width

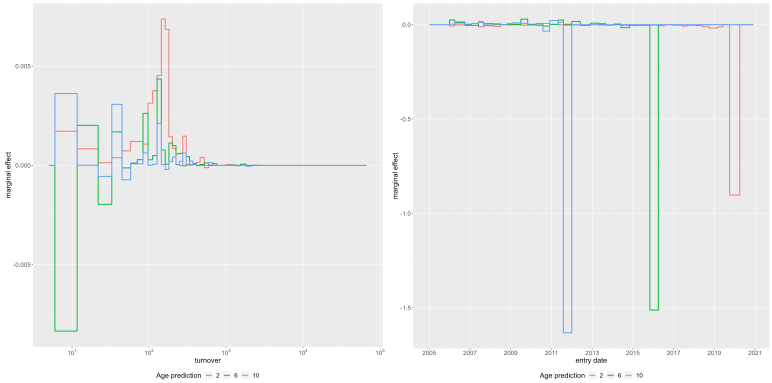
of the steps indicate the ALE-interval used to compute it.

Marginal effects in turnover are not constant but change discontinuously (see Figure 18a). At age two, the marginal effect of an increase in turnover is close to zero when turnover is below 10 thousand euros or exceeds 1 million euros. For intermediate values, the marginal effect is positive, reaching its highest value of about 0.01 when turnover is about 150 thousand euro ($10^{2.2} \approx 150$). This indicates that the probability of economic success, defined as surviving and achieving a minimum turnover of 150 thousand euros, largely depends on whether initial revenue is close to this cutoff in the first year.

The same pattern is observed for our definition of economic success at age six. However, the discontinuity is less abrupt, and the marginal effects are smaller. The impact of initial revenue on success might become smaller when firms age due to unforeseen events. In particular, the marginal effects at age ten are less clear, with marginal effects around the cutoff value of 150 thousand euros being very similar to values observed for smaller values of initial turnover. However, for all ages, the marginal effect of initial turnover on the probability of being economically successful becomes virtually zero when initial turnover exceeds a value of about 1 million euros.

In Figure 18b we present the marginal effects for firm entry dates measured as months elapsed since december 2005. The marginal effects fall strongly when firms are two, six, or ten years old in 2020, reflecting the drop in exit rates during the COVID-19 pandemic. In other years, the marginal effect of turnover on economic success later in life is quite modest.

Figure 18: Effect on economic success at various ages (20p pooled cutoff)



(a) Turnover **(b)** Entry period

2.6 What drives economic success? Capability learning vs noisy learning

The above analysis has revealed that economic success at age a, conditional on having survived up to age $a-1$, is very predictable from early life conditions, even in later years. Firm financial variables are the most important features, whose impact on the predictions cannot be taken over by other characteristics. In particular, the initial turnover level and the moment the firms enters the business environment stand out as important contributors to firm success. Next, we use our estimates to quantify the importance of dynamic selection and inspect drivers within firm dynamics. Finally, we interpret these findings using theories of noisy learning and capability learning.

2.6.1 Quantifying the importance of dynamic selection

Our rich data on firm founding conditions allows us to quantify the observable heterogeneity in firm success types at different ages. We first consider the persistence in the heterogeneity in firm success types. In

particular, we see that $R_2^2(\hat{S}_2, S_2) = 0.47$ and that $R_2^2(\hat{S}_6, S_2) = 0.38$. This suggests that about 80 percent of the observable heterogeneity in firm success types at age two is persistent over the following four years $\left(\frac{0.38}{0.47} \approx 0.8\right)$. A persistency of similar magnitude is found by Mueller and Spinnewijn (2023) for heterogeneity in unemployment risk over the first six months of the unemployment spell. More than one-third of the firm-level heterogeneity in firm success types at age two is persistent up to age ten $\left(\frac{0.17}{0.47} \approx 0.37\right)$, which is still substantial.

When considering the sample of startups that survive up to age 6, observable heterogeneity is significantly lower than in the former sample. This pattern is also found by Mueller and Spinnewijn (2023), who attribute it to selection effects within the sample. If heterogeneity is largely persistent, then, conditional on survival, the variance in success types should decrease, as startups of greater type are more likely to survive. This process leads to a monotonic decline in contemporaneous covariances, as shown in Table 11, along with a corresponding decrease in the R^2 across sample models. As a result, the impact of dynamic selection gradually diminishes over time, despite remaining a key factor in shaping success patterns. The persistence in the heterogeneity in firms' success types suggests that dynamic selection is an important driver of the increase in the firm success rate as firms grow older. This statement is reinforced by Figure 19, which illustrates the success rate at age 2 for

Table 11: Persistent heterogeneity in success types

Sample (a)	Model age (j)	N_a	$E_a[S_a]$	$E_a[\hat{S}_j]$	$Var_a[\hat{S}_j]$	$COV_a[\hat{S}_j, S_a]$	$R_a^2[\hat{S}_j, S_a]$
Age 2	2	95,738	0.6754	0.6728	0.1072	0.1053	0.4715
	6	95,738	0.6754	0.6850	0.0779	0.0811	0.3849
	10	95,738	0.6754	0.6148	0.1262	0.0694	0.1742
Age 6	6	42,095	0.9181	0.8336	0.0216	0.0049	0.0147
	10	42,095	0.9181	0.7369	0.1003	0.0033	0.0014
Age 10	10	26,802	0.9330	0.8516	0.0274	0.0100	0.0588

Notes: The sample column subsets startups in the test set based on their age a, meaning it includes only startups that have survived for at least $a - 1$ years from foundation. The model age column indicates the prediction model used. For example, $a = 2$ and $j = 2$ refer to predicting success at age 2 using startups that have survived up to age 2.

startups that have survived d years and the simple and lower bound of dynamic selection. The upper line illustrates the concave relationship between economic success and age: firms are more likely to be economically successful when they are older, yet the gain in economic success decreases with age. The second line from above provides a simple quantification of the effect of dynamic selection by taking the estimates of firm survival at age two and presenting the expectation of the estimate at every age. This line simulates a world in which survival probabilities at the firm level remain constant, and the success-age gradient fully reflects dynamic selection. In particular, we impose $E[T_{i,a}] = E[T_{i,2}]$ reflecting the theory of noisy learning. The line shows that at age three, about 88 percent of the gain in the economic success rates of firms is due to dynamic selection $\left(\frac{82.87-67.54}{85.04-67.54} = 0.88\right)$. At age five, this share has decreased to 79 percent, which is still very substantial $\left(\frac{85.83-67.54}{90.69-67.54} = 0.79\right)$. The lowest line describes the conservative lower bound to dynamic selection according to equation 2.6. The line suggests that the effects are still substantial. In this conservative view, 79 percent of the increase in firm economic success at age three is due to dynamic selection $\left(\frac{81.39-67.54}{85.04-67.54} = 0.79\right)$. This is about 55 percent at age five $\left(\frac{80.28-67.54}{90.69-67.54} = 0.53\right)$. Even at high ages, dynamic selection is important as it explains about forty percent of the increase in firm economic success between ages two and ten $\left(\frac{77.81-67.54}{93.30-67.54} = 0.40\right)$.

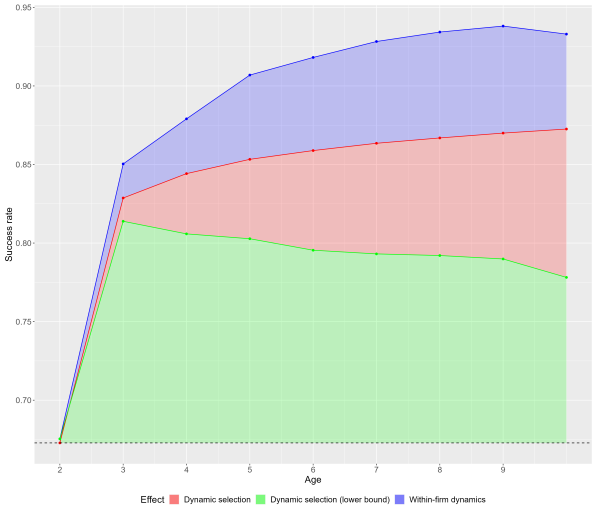


Figure 19: The blue line shows the success rate for firms given their age. The red line quantifies the role of dynamic selection as the success rate for firms that have survived up to that age when their expected success rate would not change, represented as $E_d[S_{1,XGB}]$. The green line shows the lower bound on dynamic selection described in Equation 2.6, calculated as $COV_a(S_2, \hat{S}_a)/1 - E_2[S_2]$ for $a > 2$.

Dynamic selection effects are important in all industries we consider, although the importance of dynamic selection differs across industries. In the Appendix, Figure B.5.13 presents the corresponding decomposition across various industries. The findings are broadly consistent across sectors, though the lower bound for selection effects is particularly pronounced in certain industries, such as consultancy, wholesale and retail trade, manufacturing, ICT, and agriculture. In these sectors, dynamic selection effects are especially strong during the early years, as highlighted by the high lower bound. Nonetheless, across all industries, dynamic selection plays a significant role in the initial stages, gradually diminishing as firms mature.

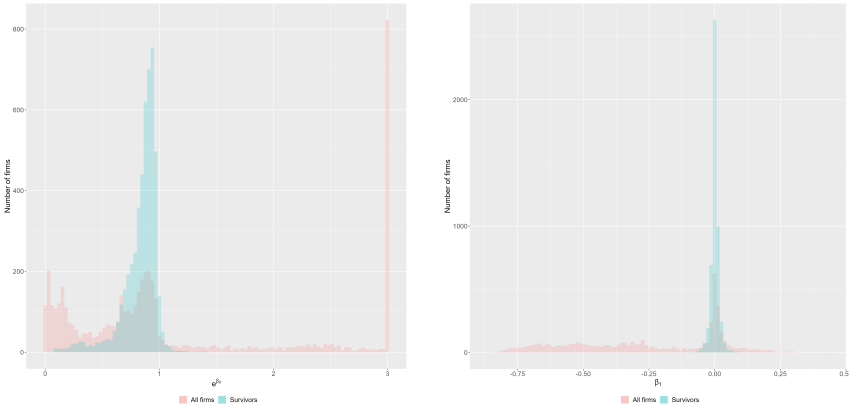
2.6.2 Within-firm dynamics

Based on the estimates of equation 2.8, we find substantial dispersion in the firm-specific effect of age on the probability of being economically successful, rejecting the hypothesis that economic success probabilities are proportional to a baseline hazard. We derive this conclusion from Figure 20, which shows the distributions of $exp(\beta_{i,0})$ and $\beta_{i,1}$. The left panel displays the distribution of the baseline hazard function, $exp(\beta_0)$ (truncated at three for about one-sixth of the sample). The distribution has two peaks, one that is concentrated around zero and one that is concentrated around 0.85. Also, the picture shows substantial overshooting of the baseline hazard function; about one-third of firms have a baseline hazard to be economically successful that is larger than one.

The distribution is very different when we consider the estimate of $exp(\beta_0)$ for firms that survive up to age ten. Then we see that the baseline hazard is distributed around one value, also about 0.85, and that it is skewed to the left. The difference in distribution indicates that initial hazard functions to be successful are larger and less dispersed for long-run survivors compared to the average firm in the sample.

The right panel depicts the expected increase in the probability of being economically successful as the firm ages by one year. We plot the distribution of this parameter for all firms and for those that survive up to age ten. Both distributions peak strongly around a value of $\beta_{i,1} = 0$. Considering all firms in the sample, the spread in the distribution to the left of zero is striking. The sample mean of $\beta_{1,i}$ equals -0.25 with a sample standard deviation of 0.28. The large standard deviation, with a bootstrapped confidence interval of 0.278 - 0.285, points to substantial heterogeneity in firm success dynamics. The average value of success dynamics is negative, suggesting that the average startup is less likely to be economically successful as it gets older. Note this is partly due to the fact that we have estimated probabilities to be economically successful for all firms in the sample, including firms that are unlikely to survive in their first years.

Figure 20: Heterogeneity in within-firm dynamics



(a) distribution of $exp(\beta_0(X_1))$ (shrunken)

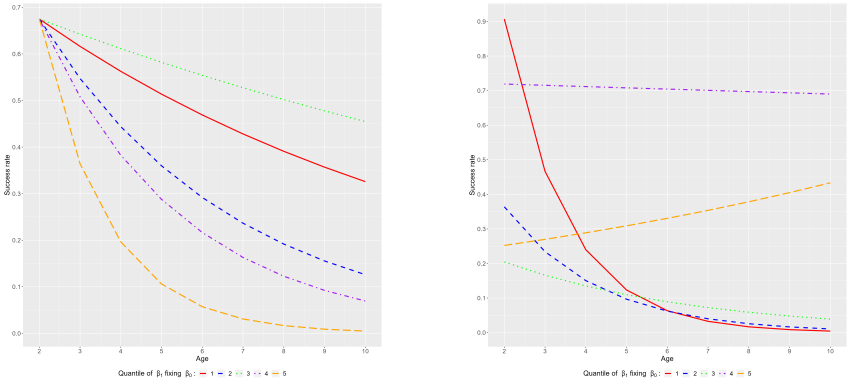
(b) distribution of $\beta_1(X_1)$ (shrunken)

In order to improve our feeling for the estimated probabilities, we plot the trajectories of $\widehat{exp(\hat{s}_{i,a})} = \hat{\beta}_{i,0} + \hat{\beta}_{i,1}a$ when firms are up to ten years old in Figure 21. In particular, we rank firms according to the (shrunken) value of $\hat{\beta}_{i,1}$, and compute the expected value by quantile. In panel 21a, we plot the development in success rate implied by these values. To improve comparison, we equate the probability of success at age two for each quantile to the sample mean (0.675). The figures show that the probability of being economically successful, which is observable early in life, decreases in the age of the firm for all quantiles, and the success trajectories are very different. Firms in the third quantile of $\hat{\beta}_{0,i}$ have the lowest decrease over time (from 0.675 to 0.45, a reduction of about one-third), but firms in the quantile have the largest (from 67.5 to 0.5, a reduction of ninety-nine percent).

Figure 21b shows the same relationship, but this time for selected quantiles of $\beta_{i,1}$. Note that success trajectories by quantiles are very different. The first quantile is successful at a young age but very unsuccessful at old age, whereas the fourth quantile is relatively successful at all ages. Note that the probability of being economically successful is increasing only for the fifth quantile. Figures 21a and 21b indicate substan-

tial heterogeneity in firm success dynamics violating the proportionality of the hazard assumption of the proportional hazard model.

Figure 21: Heterogeneity in within-firm dynamics



(a) Economic success by quantile of the baseline hazard

(b) Economic success by quantile of the age effect

Estimation results of equation 2.9 with LASSO reveal that most of the variation in linearized success rate dynamics is explained by business cycle effects. This follows from Table B.4.16, which shows that most of the selected early life conditions are year indicators.[5] The LASSO model explains about 87 percent of the variation in the linearized growth of success probability types at the firm level.

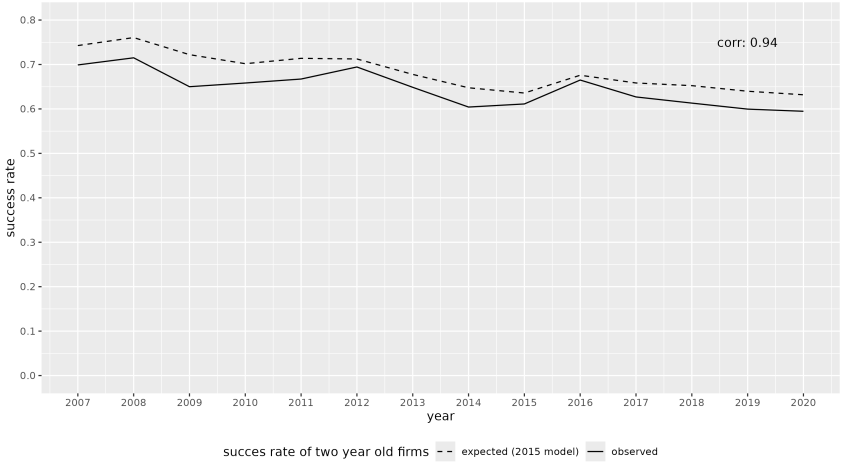
2.6.3 Heterogeneity in success rates over the business cycle

The year indicators that are important in explaining firm success dynamics could be driven by yearly changes in the composition of firms entering the market or by annual market demand shocks. Figure 22 compares the counterfactual success rate using the 2015 model to the observed

[5]When we allow for non-linearities in the relationship between the dynamics in success probabilities and early life conditions using an XGB model, we derive a similar conclusion.

success rate for the corresponding group of startups. Both series gradually decline, indicating that the share of two years old firms decreases steadily. Also, the figure illustrates that the 2015 model overestimates the share of successful firms in 2015. Despite this, the predicted and observed success rates move in the same direction as the two series are highly correlated (corr = 0.94), and changes in startup composition explain 88 percent of the annual variation in success rates of two-year-old firms. This is not specific to the 2015 model. When we repeat the analysis for other years, we find that variation in the composition of startups explains at least 86 percent and at most 92 percent of the annual variation in success rates of two-year-old firms. As we keep the prediction model fixed, this suggests that variations in the *observable* composition of firms are important for explaining annual variations in the observed success rate of young firms. Naturally, since we use a prediction model to derive the counterfactual success rate, we cannot rule out the possibility that *unobserved* characteristics fluctuate across years and affect the observed success rate of young firms. However, for this to alter our conclusion, selection based on observable heterogeneity affecting firm success would need to be fundamentally different from selection based on unobservable heterogeneity.

Figure 22: Compositional effects over the business cycle

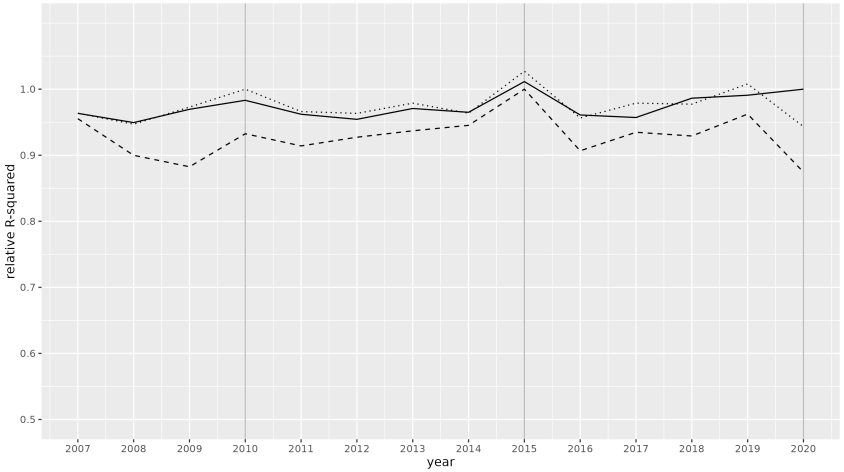


Notes: The solid line plots the share of two-year-old firms that are economically successful by year. The dashed line plots the share of firms that is expected to be successful according to the 2015 model and early life conditions of two years old firms in the years 2007 - 2020. The 2015 model is estimated using the test data for firms that were two years old in 2015.

Next, we show evidence that the relationships between early life conditions and subsequent firm success at age two are very similar over time. Figure 23 shows that the ratio of the R^2 in year t' over the baseline year t remains high throughout the sample period considered. The ratio remains about 95 to 90 percent of the optimal R^2 even when the difference in years increases to ten. Note that the relative R^2 is one by definition for the year in which the model was trained. For all years, the relative R^2 are lower than one, indicating that some predictive power is lost when the mapping from a different year than the baseline year is used to make predictions. The exception to this rule is the 2015 model, which predicts economic success in its own year relatively badly (as figure 22 showed). Also, note that the predictive power of the models remains around 90 percent during the financial crisis and the early start of the COVID-19 pandemic.

All in all, the high relative predictive performance of models trained in years that are very far away from the evaluation year suggests that early life conditions that are predictive of economic success in the one year explain firm success similarly in other years *in a similar way*. Thus, despite the fact that we document cyclicality in the economic success of young firms (Figure 22), the relationship between the early life conditions that explain economic success seems relatively stable over the business cycle. This is in line with our finding that variation in the composition of startups is the main driver for annual variation in the success rate of young firms.

Figure 23: Persistence in predictive power over time



Notes: The figure shows the R^2 in the test data of 2010, 2015, and 2020 using the prediction model for each year relative to the R^2 in the test data from the same year as the prediction model. Vertical grey solid lines indicate the years that were used in the test data.

2.6.4 Interpretation in terms of noisy learning and capability learning

According to the noisy learning hypothesis, early life conditions, such as having substantial liquidity, increase the likelihood of a firm surviving its initial years, as it provides more flexibility to absorb adverse shocks. However, early life conditions do not directly affect a firm's economic success during any given period. This is because some firms may, by chance, succeed in building a stable network of suppliers and customers, while others may fail to do so, also by chance. In particular, firms do not change behavior in response to major changes to the environment.

This contrasts with the theory of capability learning, which states that some firms are better equipped to leverage their capabilities for survival. Even if all firms have similar initial survivability at birth, this theory posits significant variation in the development of firm survival types as they age. Some firms will quickly learn how to achieve economic success, leading to a rapid increase in their survival probability, while others may fail to do so, resulting in stable or even declining survival probability over time. Capability learning assumes, therefore, variation in firm economic success dynamics that is predictable from firm characteristics.

After re-interpreting the evidence provided in Sections 2.6.1 to 2.6.3 in light of these theories, we conclude that noisy learning is an important driver of firm learning. This conclusion is derived from three observations. First, we find that success probability types of firms are persistent, which limits the degree to which capability learning takes place. Indeed, more than half of the gain in the probability of being economically successful when a firm's age increases from two to five can be explained by dynamic selection. This suggests that noisy learning is a very important mechanism by which firm learning takes place when firms are young and when firms are more mature.

Second, firm-specific changes in success rate probability types are small in size and mostly explained by year indicators. Capability learning would imply that early life conditions that vary at the firm level explain the dynamics in firm probability types, but we do not find evidence

for this. Instead, we provide evidence that the variation in the success rates of young firms over time is, to a large extent, explained by variation in early life conditions of these firms, which aligns well with noisy learning.

Third, we find that the mappings between early life conditions and the economic success of young firms are relatively persistent over time. The relationship between early life conditions and firm success at age two in year t' explains at least 90 percent of the explainable variation of this relationship in year t, even if the time gap between t and year t' grows to ten years. This suggests that firms do not respond strongly to changes in the economic environment, which is in line with the theory of noisy learning as well.

2.7 Conclusion

This chapter provides an in-depth empirical analysis of business survival and economic success based on comprehensive administrative data from the Netherlands, analyzing the whole population of newly incorporated firms in the period 2006 - 2021. We use machine learning to analyze this data and use early life conditions to predict the economic success of firms later in life. While we find no direct effect of early life conditions on economic success in later life when we control for contemporaneous features, the strong predictive power of early life characteristics arises from their correlation with firm characteristics in later life. Our findings indicate that firm success probability types exhibit substantial persistence over time. In particular, firms that are successful early on maintain their advantage as they age, while firms that struggle in the early years have a significantly higher probability of exiting.

An important empirical result is that firm success probability types are not randomly distributed, but correlate strongly over the years. The analysis using the framework of Mueller and Spinnewijn (2023) confirms that dynamic selection effects contribute strongly to the observed increase in success rates as firms age. This is because firms with a high initial probability of success continue to be successful, while weaker com-

panies are gradually filtered out. Intra-firm dynamics play a relatively minor role in explaining the rise in success rates as firms age, as we show that within-firm dynamics are predominantly driven by year indicators. This suggests that year-to-year fluctuations in the composition of startups or firms' reactions to external shocks are the main drivers of the observed success rates over time. We find that changes in the composition of firms' early life conditions explain around 88 percent of the variation in the success rate of young firms, suggesting that other explanations, such as changes in the economic environment, play a more limited role. Consistent with this, we find that the relationship between early life conditions and economic success is stable over time. These results provide strong empirical support for the noisy learning theory, which posits that initial firm conditions and market selection mechanisms largely determine long-term survival and success. The limited role of within-firm dynamics suggests that firm learning over time plays a secondary role compared to the initial sorting process. Our finding that firms' responses to the economic environment remain relatively stable over the years suggests that the scope for capability learning to drive success dynamics is limited. Companies can adapt, but these adaptations do not appear to fundamentally alter the observed success type probabilities of firms.

Despite the robustness of these findings, some limitations remain. Not all aspects of capability learning are directly observed in the study, making it difficult to disentangle the exact mechanisms that drive firm adaptation over time. Future research should incorporate direct measures of firm investment in innovation, managerial abilities, and strategic flexibility to enhance the interpretation of learning dynamics. Although the analysis successfully captures the long-term persistence of firm success types, policy variations could further refine the explanatory power of the model.

Overall, this research provides strong empirical evidence that firm success probability types are persistent over the life of the firm and across years, reinforcing the notion that firm trajectories are shaped early in their life course. However, the interplay between selection mechanisms and within-firm learning dynamics remains an area for further explo-

ration. By improving our understanding of these factors, future studies can inform more targeted policy interventions. Given that firm trajectories are largely determined early in their life course, policies aimed at improving initial firm conditions—such as better access to finance, mentorship programs, and early-stage business support—may have long-lasting impacts on firm success rates.

Chapter 3

Innovation and Firm Growth: A Machine-Learning Enhanced Tobit Model

The chapter is co-authored with Massimo Riccaboni.

Disclaimer: AI has been employed in the preparation of this chapter, specifically for grammar and language refinement. All content, ideas, and interpretations remain the original work of the author and coauthors.

3.1 Introduction

Firm growth dynamics represent a cornerstone of economic research, with particular attention paid to factors that drive growth—most notably innovation—and the validity of Gibrat's law, which posits that firm growth is independent of its size (Gibrat, 1931; Sutton, 1997; Kalecki, 1945). However, empirical investigations are often hindered by sample attrition issues, as first evidenced by B. H. Hall (1986) and Evans (1987b) and Evans (1987a). This estimates may conceal the true relationship be-

tween growth and its determinants, such as size, intellectual property, and other firm-specific characteristics. Traditional approaches, such as the Tobit 2 model, often address this challenge with Heckman's estimator (Heckman, 1979): a Probit model for the first stage to capture the selection mechanism, followed by a regression for the second stage to predict the outcome while correcting for selection bias through the Inverse Mills ratio. While the model is relatively straightforward to estimate under its assumptions, it suffers from significant limitations due to its parametric nature, which restricts its ability to capture complex patterns in firm observability (S. Athey, 2017). Extensions that preserve the simplicity of the estimation procedure, while substantially enhancing the reliability and interpretability of the final estimates, can provide a useful tool with practical relevance for both scholars and practitioners. Traditional methods for correcting selection bias are primarily designed to address exit attrition, which arise because growth can only be measured for firms that survive throughout the study period. However, depending on the data source, unobserved growth may result from additional factors that require careful consideration. Beyond exit attrition, growth is often observable only for a subset of firms or time periods due to reporting thresholds or data limitations. Unobserved growth is frequently accompanied by missing values in other key variables, further complicating the analysis. This issue is particularly evident in data from Orbis Bureau Van Dijk (Kalemli-Ozcan et al., 2015; Gal, 2013), where the quality of data varies widely depending on the country, industry, and size of the firm. Missing data often occur due to firm exits, regulatory exemptions, or inconsistent reporting practices, especially among smaller or privately held firms (Bajgar et al., 2020; Glaeser and Omartian, 2022; Kalemli-Özcan, Sørensen, et al., 2024). Such gaps in data not only present technical challenges but may also introduce additional biases, as the missingness itself can be informative, reflecting underlying latent dynamics. Addressing both growth censoring and the missingness in the predictors is therefore essential for producing accurate and robust estimates of firm growth dynamics.

In this study, we introduce and test an extension to Heckman's two-

step procedure, which preserves the Tobit 2 model's traditional assumption of joint normality of errors while replacing the Probit model in the first stage with a machine-learning algorithm. By leveraging the ability of machine-learning models to capture complex non-linear relationships and interactions among predictors, this approach provides a more robust and flexible method for predicting the likelihood of observing firm growth rates (Mullainathan and Spiess, 2017). In the second stage, the machine-learning-based Inverse Mills ratio is incorporated into the growth rate prediction model. This enhances the model's ability to account for the censoring mechanism and correct for selection bias. Our application focuses on predicting firm-level growth while addressing selection effects. We explore how the relationship between growth, size, and innovation changes according to the controlled selection mechanism. In particular, we examine the role of intellectual property—such as patents and trademarks—as potential drivers of firm growth. Trademarks and patents are key indicators of innovation and strategic differentiation, and their effects on firm growth dynamics are of significant interest in both theoretical and applied research (B. Hall et al., 2014; Greenhalgh and Rogers, 2010).

Our contribution is threefold. First, we address the methodological limitations of traditional traditional Heckman's procedure in Tobit 2 models by incorporating machine-learning tools, showing their superior ability to effectively capture the censoring mechanism and correct for potential selection biases. Second, we empirically assess whether the common rejection of Gibrat's law depends on the choice of model, providing insights into its robustness when applied to widely used firm-level data sources. Third, we investigate the role of innovation in shaping firm growth trajectories. Our findings indicate that the rejection of Gibrat's law is largely independent of the selection mechanism, reaffirming the statistical regularity of the negative relationship between firm size and growth. However, we find that innovation have a positive impact on growth only when machine learning is employed to enhance Heckman's model. This shows the value of integrating machine learning to address censoring and missing data, providing a more robust framework for an-

alyzing firm growth dynamics.

3.2 The conceptual framework

According to the seminal work of Heckman (1979), the bivariate sample selection model also known as Tobit 2 model, specifies: (i) a model for the censoring mechanism; (ii) a model for the outcome, conditional on the outcome to be observed. In our application, the Tobit 2 model consists of two main components: a participation equation that determines the observability of the growth rate and an outcome equation that determines its value. We will use throughout the paper the notation employed in Cameron (2005). The participation equation, which describes the censoring mechanism of the growth rate, is given by

$$y_1 = \begin{cases} 1 & \text{if } y_1^* > 0 \\ 0 & \text{if } y_1^* \leq 0 \end{cases}$$

where y_1^* is a latent variable. The outcome equation, which describes the value of the growth rate, is

$$y_2 = \begin{cases} y_2^* & \text{if } y_1^* > 0 \\ - & \text{if } y_1^* \leq 0 \end{cases}$$

The latent variables y_1^* and y_2^* are linear combinations of explanatory variables plus an error term

$$\begin{aligned} y_1^* &= \mathbf{x}_1' \boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2^* &= \mathbf{x}_2' \boldsymbol{\beta}_2 + \varepsilon_2 \end{aligned} \tag{3.1}$$

Under the assumption that the errors are jointly normally distributed, homoskedastic and correlated, the model can be estimated through maximum likelihood estimation (MLE). Under this assumption in the model, analytical computations based on the properties of the truncated moments of the normal distribution show that the truncated mean can be

expressed as

$$\mathbb{E}\left[y_2 \mid \mathbf{x_1}, \mathbf{x_2}, y_1^* > 0\right] = \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)$$

where $\lambda(z) = \phi(z)/\Phi(z)$ is the Inverse Mills ratio (IMR), defined as the ratio between the density $\phi(z)$ and the cumulative $\Phi(z)$ distributions of the standard normal. Therefore, simply projecting y_2 onto the space spanned by $\mathbf{x}_2$ results in a biased estimate of $\boldsymbol{\beta}_2$ unless a correction for the sample selection mechanism is applied.

While the traditional MLE approach is more efficient under the joint normality assumption of the error terms, in our setting the computational burden is high, as well as sensitive to potential misspecifications in the underlying model. The Heckman estimator provides consistent estimates by explicitly separating the selection process from the outcome equation through the incorporation of the IMR, making it a robust and practical choice for model estimation. The Heckman procedure augments the second-stage OLS regression with an estimate of the omitted regressor $\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)$. Using the observed values of y_2, the model is estimated as

$$y_2 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\lambda\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right) + v$$

where v is an error term. The parameter $\boldsymbol{\beta}_1$ is obtained from a first-step probit regression of y_1 on $\mathbf{x}_1$. Specifically, the probability of $y_1 = 1$ is given by

$$\Pr[y_1 = 1|\mathbf{x}_1] = \Phi\left(\mathbf{x}_1'\boldsymbol{\beta}_1\right)$$

In our application, we propose an alternative approach where we replace the traditional probit model used in the first step of the Heckman procedure with predictions from a machine learning model. It is worth noting that the Heckman model's assumption of joint normality of the error terms is preserved in our extension. What we boost, in essence, are the selection probabilities from the first stage. In our case, these probabilities are generated by a more complex non-linear model, which leads to improved predictions in the first step of the two-stage estimation procedure. The core idea is to leverage the flexibility and predictive power of machine learning algorithms, which can better handle complex re-

lationships and interactions in the data compared to the standard probit model. The first step involves training a machine learning model to predict the probability of observing the growth rate. Once the machine learning model is trained, we obtain predicted probabilities for each observation. Next, we convert these predicted probabilities into the probit scale by applying the inverse of the cumulative distribution function of the standard normal distribution, denoted as $\Phi^{-1}(z)$. This transformation maps the predicted probabilities $\hat{p}_1$=Pr$[y_1 = 1|\mathbf{x}_1]$ onto the latent variable scale, which corresponds to $\mathbf{x}_1'\boldsymbol{\beta}_1$ in the traditional probit model. Then, we compute the IMR as follows:

$$\hat{\lambda}_{\text{ML}} = \frac{\phi(\Phi^{-1}(\hat{p}_1))}{\Phi(\Phi^{-1}(\hat{p}_1))} = \frac{\phi(\Phi^{-1}(\hat{p}_1))}{\hat{p}_1}$$

where ML stands for machine-learning. In the second stage, we then perform the outcome regression using the machine learning-based IMR. The outcome equation becomes

$$y_2 = \mathbf{x}_2'\boldsymbol{\beta}_2 + \sigma_{12}\hat{\lambda}_{\text{ML}} + v$$

where v is the error term and $\hat{\lambda}_{\text{ML}}$ is the estimated IMR based on the machine learning predictions.

3.3 A case for missing data

In the Heckman settings, where the first stage predicts the likelihood of observing the growth rate, machine-learning tools offer several advantages over traditional probit models. This is particularly important in empirical analyses where predictors often contain missing values. Unlike standard probit models, which typically require Complete-Case (CC) analysis or imputation, some machine-learning techniques are specifically designed to handle missing values directly. This capability enables the extraction of meaningful patterns without the need for preprocessing steps that could reduce the sample size or, worse, introduce biases—especially in the case of Missing Not At Random (MNAR) data

(Little and Rubin, 2019), where missing values can provide important information for prediction. Furthermore, machine-learning models are outstanding to capture complex, non-linear relationships and interactions among predictors, offering superior predictive accuracy in scenarios where traditional parametric models fall short. As a result, using machine-learning approaches in the first stage of a Heckman procedure not only simplifies data preparation but also enhances the robustness and flexibility of the estimation process, especially when dealing with censored growth rates and incomplete predictors in our data.

Among the many machine-learning tools available, eXtreme Gradient Boosting (XGBoost) stands out as a particularly effective model for this purpose. It is an ensemble method based on decision trees within a gradient boosting framework, renowned for its high performance across diverse prediction tasks. Its computational efficiency in terms of time and memory makes it a reliable choice for large datasets (Gumus and Kiran, 2017; Abbasi et al., 2019; M. Li, Fu, and D. Li, 2020). A key strength of XGBoost is its ability to handle missing values through the so-called sparsity-aware algorithm (T. Chen and Guestrin, 2016). This technique directs missing observations for continuous variables to the branch that minimizes error, ensuring effective use of incomplete data. In contrast, categorical variables require one-hot encoding, with missing values treated as a separate category during preprocessing. These mechanisms, combined with regularization techniques to prevent overfitting, make XGBoost particularly well-suited for datasets with missing and complex features (Bentéjac, Csörgő, and Martínez-Muñoz, 2021).

The utility of XGBoost becomes especially apparent when dealing with datasets such as Orbis, which present significant challenges due to missing values. Missing data in Orbis can occur for various reasons, such as firm failure, changes in ownership, or operational status (Kalemli-Özcan, Sørensen, et al., 2024; Liu, 2020). Additionally, some firms may strategically restrict financial disclosures or simplify their records, as they are not obligated to disclose or even calculate certain financial variables (Falco J Bargagli-Stoffi, Incerti, et al., 2024b). For example, growth rates may be unavailable when firms exit the market or fail to provide consistent

financial statements during key periods. These issues are particularly pronounced in private firms, which face fewer regulatory reporting requirements than publicly listed firms (Glaeser and Omartian, 2022). Furthermore, Orbis data quality varies across countries, industries, and firm sizes, with SMEs often underrepresented (Dinlersoz et al., 2018). In this context, XGBoost's ability to process incomplete information robustly ensures that first-stage estimations in Tobit 2 models remain both unbiased and efficient, making it a valuable tool for addressing the complexities of real-world datasets.

3.4 Data

Our work relies on firm-level data sourced from the ORBIS database, a popular repository for academics and practitioners of global financial accounts provided by Bureau Van Dijk. ORBIS data have been extensively applied in prior economic research (Gopinath et al., 2017; Cravino and Andrei A Levchenko, 2017; Adalet McGowan, Andrews, and Millot, 2018; Gourinchas et al., 2020; Kalemli-Özcan, Laeven, and D. Moreno, 2022). Despite its comprehensive nature, ORBIS coverage can vary across countries due to differences in national business registry filing requirements (Gal, 2013; Kalemli-Ozcan et al., 2015). For Italy, financial account data are primarily collected from CERVED, a credit rating agency, before being standardized and translated by Bureau Van Dijk to facilitate cross-country comparability. Additionally, ORBIS enriches financial records by integrating supplemental information on firm ownership, management structures, and intellectual property rights, which prove valuable for predictive modeling. This study employs ORBIS data for Italian manufacturing firms active for at least one year between 2008 and 2017 (see Falco J Bargagli-Stoffi, Incerti, et al., 2024b for additional information about the data). We first compute the firm-level proportional growth:

$$\pi_n = \frac{size_{t+n} - size_t}{size_t} \tag{3.2}$$

and we define the growth rate according to Capasso, Treibich, and Verspagen (2015):

$$g_n = log\left(\pi_n + 2\right) \tag{3.3}$$

where $size_t$ represents firm's revenues (or employment) at year t and n is the growth window. The transformation strikes a balance between interpretability and robustness. For firms experiencing high growth rates, it closely approximates the log difference growth, preserving the relationship between size changes and growth. However, unlike the proportional measure, it reduces the impact of heteroscedasticity in growth, as empirical evidence shows that growth variance is dependent on firm size (B. H. Hall, 1986; Evans, 1987b; Evans, 1987a). That is because smaller firms tend to exhibit higher variability in growth rates compared to larger firms. Moreover, in cases of extreme negative growth, the measure is also less sensitive to endogenous attrition. In our analysis, we focus on a three-year period ($n = 3$), with additional robustness checks for one and two-year periods provided in the Appendix. Our dataset includes a wide range of economic and financial indicators, such as original financial accounts, measures of financial constraints, and indicators of financial risk. The full list of predictors is available in Falco J Bargagli-Stoffi, Incerti, et al. (2024b), as we use the same data. Additionally, the dataset includes dummy variables for trademarks (equal to 1 if the firm has issued trademarks and 0 otherwise), patents (equal to 1 if the firm has issued patents and 0 otherwise), and consolidated accounts (equal to 1 if the firm consolidates the accounts of its subsidiaries). However, in Tobit 2 models, incorporating a large set of predictors, such as three years of lagged variables, can lead to challenges such as multicollinearity, where high correlations among predictors inflate standard errors and reduce the precision of parameter estimates. To address this, we perform a dimensionality reduction through Principal Components Analysis (PCA). This way, we transform the original predictors into a smaller set of uncorrelated components that capture the majority of the variance in the data, mitigating multicollinearity while preserving the most relevant information. Additionally, reducing the dimensionality of the predictor space improves computational efficiency and helps prevent overfitting.

By simplifying the predictor set, PCA enables a more robust estimation process and ensures that the model remains interpretable and statistically sound, even in the presence of a large and complex dataset. In the data construction pipeline, we begin by calculating the growth rate over a three-year period. Next, we split the data, allocating 80% to the training set and 20% to the test set. The training set is used to train both the first and second stage regressions, while the test set is employed for evaluating the first stage model's performance. Since XGBoost can handle missing values in the predictors, it can be applied to a larger set of observations compared to traditional models, which require CC data. To ensure fairness, we construct the training and test sets for XGBoost using a random sample of the same size as the CC data used in traditional methods, while allowing observations with missing predictors to be included.

3.5 Results

3.5.1 Well-predictable missing growth

We first analyze the ability to predict the missing growth rate, focusing on revenue-based growth and later assessing its link to innovation outcomes. However, we also find it valuable to compare it with employment-based growth, as traditional literature primarily emphasizes the latter. Employment growth represents an input, while revenues growth reflects an output, offering distinct perspectives on firm performance (Coad, 2009). For example, innovation may allow firms to achieve higher outputs with fewer inputs, potentially increasing revenues without a corresponding rise in employment. This distinction is important since policymakers are often concerned with employment trends and their implications, whereas firms are more concerned about revenue and profit growth as key indicators of success and competitiveness. Comparing both measures provides a more comprehensive understanding of firm dynamics.
Table 12 compares the performance of probit, random forest, and XGBoost in predicting the probability of missing growth, which provides

the first stage of Heckman's estimation. The AUC-ROC scores, which measure the model's ability to distinguish between firms with observed and unobserved growth, show that the probit model has limited performance for both classes of growth compared to its competitors, reflecting a limited ability to capture complex patterns. Random forest achieves higher scores, indicating stronger predictive ability, but XGBoost shows consistently superior performance, achieving near-perfect AUC-ROC scores (0.9807 and 0.9879, respectively), underscoring its robustness and ability to accurately differentiate between missing and non-missing growth. Furthermore, we assess predictive performance using the AUC-PR (Area Under the Precision-Recall Curve). Although our data exhibit only a moderate imbalance (52% of revenue growth is missing and 60% for employment), the AUC-PR underscores the superior performance of XGBoost in terms of both precision and recall. Likewise, the F1-Score, which balances precision and recall, highlights the relative strengths of the machine learning models. Similarly, the R^2, which measures the proportion of variance explained by the model, reveals that XGBoost captures data patterns most effectively, with random forest trailing behind. Finally, the Balanced Accuracy (BACC), reflecting the average of sensitivity and specificity, further underscores XGBoost's superiority. Thus, the superior predictive power of XGBoost, particularly its robustness, precision, and ability to exploit missing patterns effectively, makes it a suitable choice for the first-stage estimation in this setting. Importantly, all metrics are computed on the test set, allowing us to assess the generalization ability of the model out-of-sample. The high performance across all evaluation criteria suggests that the model does not suffer from overfitting, reinforcing the reliability of its predictions in capturing the selection mechanism.

3.5.2 Unpacking first stage modelling

The first stage estimation of the missing growth is based on features measured at the beginning of the growth window, along with other time-invariant controls. Despite the limited information available to the algorithm, the strong performance of XGBoost is likely depends on its ability

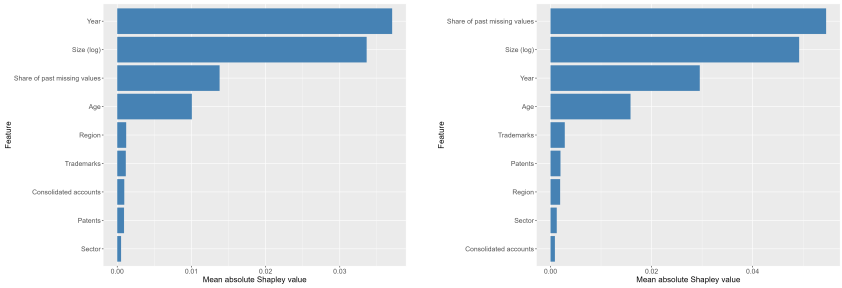
Table 12: First stage out-of-sample prediction (missing g_3)

Method	AUC-ROC	AUC-PR	F1-Score	R^2	BACC	Num. obs
(a) Employment						
Probit	0.6862	0.2109	0.2434	0.0499	0.6293	82,096
Random forest	0.7402	0.2541	0.2901	0.0837	0.6790	82,096
XGBoost	0.9807	0.9899	0.9625	0.8517	0.9638	82,096
(b) Revenues						
Probit	0.8271	0.4966	0.3336	0.2755	0.7449	92,920
Random forest	0.8479	0.5217	0.3572	0.3023	0.7665	92,920
XGBoost	0.9879	0.9918	0.9702	0.8941	0.9709	92,920

to capture non-linearities that probit models fail to detect. Moreover, missing values themselves may also contain valuable information that both probit and random forest cannot exploit because of their CC nature. However, assessing the importance and quality of the signal provided by missing values is challenging. In that respect, we introduce in the first stage a missingness indicator, which provides the share of missing values in financial statements features over the two years prior the growth window. To assess variable importance, we compute approximate Shapley values for the XGBoost model. Given the high computational burden, we use a random sample of 15,000 observations. Figure 24 shows the mean absolute Shapley values in both revenues and employment-based three-year growth models. In both models, the share of past missing values, size, and year are the most important predictors. This suggests that past patterns of missing data are highly informative in the prediction. Furthermore, due to the XGBoost use of default directions when splitting on a feature—which assigns all missing values to either the left or right branch—we expect missing values in size to play a critical role in generating missing growth in subsequent periods. Consequently, size is likely to affect the prediction through two main channels: its missing values and its observed values. This further supports that attrition is often driven by Orbis-specific data collection dynamics. Additionally, the high importance of year aligns with differences in Orbis data coverage over time (Kalemli-Ozcan et al., 2015), which may lead to higher

rates of missing growth in specific years. Figure C.2.1 in the Appendix shows the Shapley values, using an equivalent CC random sample, for the probit model. Interestingly, the probit relies on the same predictors as XGBoost. This suggests that the higher ability to separate missing and non-missing growth is likely driven by the signal from missing values and non-linearities captured by the tree-based nature of XGBoost.

Figure 24: Shapley values for XGBoost first stage prediction (missing g_3)

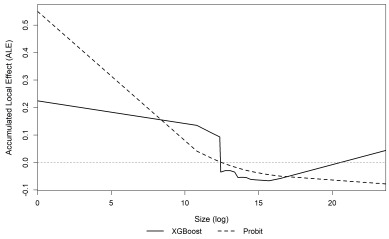


(a) Revenues **(b)** Employment

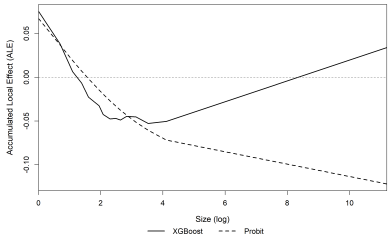
To further investigate how (observed) size affects the probability of missing growth, we compute and compare Accumulated Local Effects (ALE) plots. Figure 25 shows them for size and share of past missing by growth model. For both models, the probit exhibits a monotonic downward trend, indicating that larger firms are less likely to experience unobserved growth. This suggests that, in general, firm size is negatively correlated with the probability of missing growth. However, the XGBoost model shows a more nuanced relationship. In the case of employment growth, the ALE initially decreases but reverses for larger firms. This implies that after reaching a certain size threshold, the probability of missing employment growth begins to increase again. Interestingly, both algorithms agree that firms with revenues exceeding $270,000$ ($\approx e^{12.5}$) euros contribute negatively to the probability of unobserved growth. Beyond this threshold, the XGBoost model shows that the contribution levels off and slightly increases again for very large firms, suggesting that

very large firms are more likely to exhibit missing growth. This finding indicates that while larger firms generally have more reliable revenues growth, there may be discontinuities in the data where revenues growth is unobserved for some large firm. Regarding, the impact of past missing values, the probit shows an almost linear and negative effect on the probability of missing growth. Interestingly, both algorithms converge on the finding that as the share of missing values over the past two years approaches one, the probability of unobserved growth decreases. However, XGBoost spots more complex patterns, particularly at lower quantiles. In the revenue-based model, the relationship appears bimodal, with two peaks in the probability of missing growth—one when the share of past missing values is low (around 0.2) and another as it approaches one. The latter may reflect cases where firms are already registered in Orbis, but the actual information is systematically processed or updated later. This lag likely explains this trend. Similarly, in the employment-based model, XGBoost captures a plateau in the accumulated local effect when the share of missing values is at most 0.5, followed by a sharp decline as missingness increases further. This suggests that missing data may not provide uniform signals but could also reflect idiosyncratic firm-level reporting and data collection behaviors, which traditional linear models like probit fail to capture.

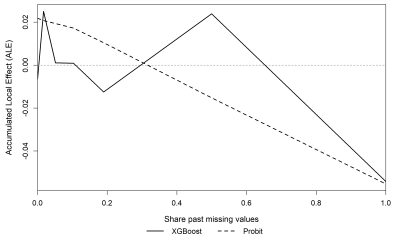
Figure 25: ALE plots for size



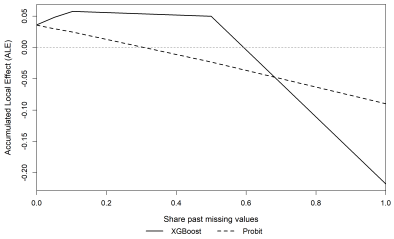
(a) Size (Revenues)



(b) (Size) Employment



(c) Share past missing (Revenues)



(d) Share past missing (Employment)

3.5.3 Outcome model

Table 13 reports the second-stage regression results for estimating the growth rate over a three-year period, comparing the naive model (without correction for censoring) with traditional (probit) and extended (XGBoost) Heckman's models. While the definition of the growth rate in equation 3.3 partially mitigates the issue of heteroscedasticity, evidence of heteroscedasticity remains present. To address it and ensure more robust inference, we employ heteroscedasticity-consistent standard errors using the HC3 estimator (MacKinnon and H. White, 1985), which is par-

ticularly suited for handling influential observations[1]. The coefficient for firm size (logarithmically measured at the start of the growth period) is consistently negative across all models. However, a strong rejection of Gibrat's law is observed only in the employment-based models. In these models, the relationship between growth and size keeps robust regardless of the inclusion and design of sample selection corrections. This divergence between models is not surprising, as employment-based growth reflects changes in workforce and operational capacity. In contrast, revenue-based growth can be influenced by external factors such as market prices or sales strategies, which may weaken the observed relationship between size and growth. Innovation indicators (patents and trademarks) exhibit nuanced effects on growth. In the naive model (a), both patents and trademarks show positive and significant effects on both revenues and employment growth. However, the inclusion of selection mechanisms changes the significance and direction of these coefficients. Notably, the probit-based model (b) shows insignificant coefficients for patents (-0.027) and trademarks (0.002) in revenues models. By contrast, the extended model (c) shows positive effects of innovation, with patents and trademarks once again emerging as significant drivers of both revenue and employment growth. These findings align more closely with the prevailing theoretical and empirical literature on innovation. Additionally, firm age is consistently negatively associated with growth across all models, which is also in line with earlier findings. The magnitude of this negative relationship is modest but highly significant in all specifications, indicating that younger firms tend to grow faster. Consolidated accounts have a consistently strong and positive effect on growth across all models and outcomes, with coefficients particularly high in the probit-based models (2.711 and 1.936 respectively). The IMR provides critical insights into the underlying selection dynam-

[1]The HC3 estimator is given by:

$$\widehat{\text{Var}}(\hat{\beta}) = (X^\top X)^{-1} \left(\sum_{i=1}^{n} \frac{\hat{u}_i^2}{(1-h_{ii})^3} x_i x_i^\top \right) (X^\top X)^{-1}$$

where $\hat{u}_i$ is the residual for observation i, and h_{ii} is the leverage for observation i (i.e., the diagonal of the hat matrix $H = X(X^\top X)^{-1} X^\top$).

ics captured by the different models. In the probit-based model (b), the IMR is positive and highly significant (1.939 for revenues growth and 0.202 for employment growth), providing strong evidence of positive selection bias: firms with observable growth are more likely to exhibit higher growth rates. In contrast, the XGBoost-based model (c) shows a negligible and statistically insignificant IMR for revenues growth (0.004). The model thus suggests that the selection bias requiring correction is minimal. Given XGBoost's superior ability to fit the first stage, it is likely more effective at identifying selection patterns related to the observability of growth rates. However, these dynamics introduce little bias in the estimation of the outcome model. Interestingly, the IMR for employment growth in model (c) is negative and significant (-0.012), suggesting a slight negative selection bias, where firms more likely to be observed tend to exhibit lower growth. In this scenario, selection occurs with smaller and younger firms, which typically exhibit higher growth but have more limited coverage in financial and employment data compared to the firm and Orbis population (Bajgar et al., 2020). On the revenues growth side, the selection process may be influenced by more random data collection dynamics within Orbis, as indicated by the insignificant IMR in the model.

Table 13: Second-stage regressions: three-year models comparison

Three-year growth rate (g_3)						
	Revenues			Employees		
	(a)	(b)	(c)	(a)	(b)	(c)
Intercept	-59.522^{***}	-1826.469^{**}	-65.553^{***}	-25.169^{***}	-817.613^{***}	-16.571^{***}
	(2.040)	(605.971)	(4.477)	(1.721)	(49.928)	(2.091)
Age	-0.001^{***}	-0.010^{**}	-0.001^{***}	-0.001^{***}	-0.002^{***}	-0.001^{***}
	(0.000)	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)
Size (log)	-0.005^{*}	-0.396^{**}	-0.005^{*}	-0.030^{***}	-0.410^{***}	-0.027^{***}
	(0.002)	(0.136)	(0.002)	(0.001)	(0.024)	(0.001)
Consolidated accounts	0.073^{***}	2.711^{**}	0.077^{***}	0.095^{***}	1.936^{***}	0.087^{***}
	(0.014)	(0.912)	(0.015)	(0.010)	(0.121)	(0.010)
Patents	0.026^{***}	-0.027	0.026^{***}	0.024^{***}	-0.199^{***}	0.026^{***}
	(0.002)	(0.018)	(0.002)	(0.002)	(0.014)	(0.002)
Trademarks	0.019^{***}	0.002	0.018^{***}	0.027^{***}	-0.228^{***}	0.029^{***}
	(0.003)	(0.006)	(0.003)	(0.002)	(0.016)	(0.002)
IMR		1.939^{**}	0.004		0.202^{**}	-0.012^{***}
		(0.665)	(0.003)		(0.077)	(0.002)
PCA Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
First Stage		Probit	XGB		Probit	XGB
Second Stage	OLS	OLS	OLS	OLS	OLS	OLS
R^2	0.0414	0.0563	0.0414	0.0470	0.0563	0.0477
RMSE	0.2461	0.1955	0.2461	0.1964	0.1955	0.1963
Obs. 1st Stage		371,028	371,028		325,165	325,165
Obs. 2nd Stage	82,861	82,861	82,861	79,026	79,026	79,026

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: PCA controls include lagged predictors condensed into five principal components. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent (HC3) estimation.

In the appendix, we provide robustness checks for one and two year growth analyses. In the first-stage predictions, XGBoost consistently outperforms both probit and random forest across all periods, showing its robustness when predicting the observability of growth across different horizons. Tables C.1.4 and C.1.5 highlight the challenges of predicting short-term growth (one and two years ahead), where shorter growth windows result in lower R-squared and RMSE. This reflects the high volatility and dominance of transitory factors in short-term growth, which reduce the model's ability to capture systematic trends (Coad, 2009; Haltiwanger, Jarmin, and Miranda, 2011). Increasing R-squared with longer

growth windows reflects the model's ability to capture more systematic variation as transient noise diminishes over time. However, RMSE also increases because longer horizons amplify the scale of growth variability, leading to larger absolute prediction errors. The negative effects of size and age remain significant for employment-based growth, whereas this relationship is weaker for the revenue-based one. Innovation effects on growth are negative or insignificant in the probit-based model, further supporting our claim that traditional sample selection corrections may be inadequate, even in the short run.

In Table C.1.6, we provide additional robustness checks using a random forest model in the first stage, yielding results very similar to those obtained with XGBoost. This shows that, despite its ability to capture informative missingness in predictions, random forest can still provide effective predictions and corrections for Heckman's model. However, the computational burden is significantly higher than XGBoost. Considering its ability to handle missing values, shorter runtime, and strong predictive performance, XGBoost is a suitable choice for the first stage. We also generalize the first stage to account for the joint probability that a firm has missing growth and is excluded from the second stage due to missing predictors. Table C.1.7 shows similar results to those from the traditional first stage. To further investigate the effects of firm size, in Table C.1.8, we conduct a panel regression with individual and time-fixed effects, allowing for better control of time-invariant unobserved heterogeneity. However, this approach comes with the trade-off of excluding time-invariant variables, such as patents, trademarks, and consolidation, due to collinearity issues. We find that the negative relationship between size and growth is stronger for employment-based growth and becomes more pronounced as the growth window increases. The magnitude of this effect is notably larger, indicating that, after accounting for time-invariant confounders, younger firms exhibit significantly higher growth rates than older firms, and this result holds regardless of the sample selection corrections applied.

3.6 Limitations and future research avenues

In this application, we aim to highlight how the selection process underlying the dynamics of firm growth is significantly more complex compared to the design of traditional empirical models. Specifically, widely-used firm-level data available to researchers and practitioners often shows limitations that must be carefully considered when drawing inferences about firm dynamics. At the same time, our contribution has several methodological limitations. In particular, by employing gradient boosting in the first stage, our model implicitly assumes that the observability of the growth rate is not driven by a latent variable that linearly depends on the features of interest (i.e., $\mathbf{x}_1'\boldsymbol{\beta}_1$), as in the standard Heckman framework. Instead, we allow for a nonlinear relationship of the form $f(\mathbf{x}_1;\boldsymbol{\gamma}_1)$, where $\boldsymbol{\gamma}_1$ represents the parameters uniquely identifying the model. Although the model provides more flexibility in capturing the selection process, we keep the assumption of normally distributed error. The two-step Heckman (heckit) procedure should rely on MLE of $\boldsymbol{\gamma}_1$, as for $\boldsymbol{\beta}_1$ when estimating the probit model. However, this estimation process is inherently complex due to the nonparametric nature of gradient boosting. To tackle this challenge, we instead use a two-step approach. First, we estimate the selection model using gradient boosting. Secondly, we incorporate the predictions into the outcome equation through the IMR. While this "backward blending" approach allows us to estimate the model, it also highligths the methodological limitations in extending the first-stage specification more rigorously. Additionally, the support of the IMR derived from our machine learning approach is more restricted compared to the traditional model. This is because, in the traditional approach, the IMR is based on the extrapolation of $\mathbf{x}_1'\boldsymbol{\beta}_1$ from the probit estimation. In contrast, in our case, the corresponding IMR is based on the extrapolation of gradient boosting predictions, which rely on a binary label. As a result, their support is concentrated around the $[0, 1]$ region. While using an appropriate loss function can extend this support over the boundaries, the estimates still revolve around the interval. Consequently, the lower variability of the IMR may explain the reduced

statistical significance of the corresponding coefficients, as well as, the higher R^2 in the second stage when the probit model is employed to correct for selection. An interesting avenue for future research lies in the use of Bayesian methods (O'Neill, 2024) to flexibly handle selection and censoring issues with a fully nonparametric approach. Such advancements could provide more flexible and robust tools for modeling the observability mechanisms underlying firm growth dynamics.

Our approach is conceptually related to double machine learning (Chernozhukov et al., 2018). In particular, the use of machine learning models to predict the selection mechanism can be interpreted as a way to partial out nuisance components from the second-stage structural equation of interest—namely, the outcome equation in our framework. This aligns with the double-machine learning approach, which stresses orthogonalization. Although the orthogonality conditions are not explicitly imposed in our model, the two-stage structure—where machine learning is applied flexibly in the first stage and the Inverse Mills Ratio is used to correct for selection in the second—mirrors the spirit of double machine learning by mitigating overfitting and model misspecification in the selection process. In that respect, it is worth noting that Hirukawa et al. (2023) introduce the DS-HECK method based on double-lasso estimation for Heckman selection models. This approach is particularly useful when the first-stage model involves a high-dimensional and sparse set of predictors, potentially exceeding the sample size. In contrast, our application primarily relies on theoretical-based predictors of missing growth, such as the current size and past patterns of missingness. While the DS-HECK framework offers valuable tools when facing a large number of potential predictors, this is not the focus of our study. Our primary aim is to capture nonlinearities and the impact of missing data on prediction, contexts where high-dimensional sparsity is less relevant. Nevertheless, the methodological innovations introduced by DS-HECK provide a complementary perspective, especially in applications where variable selection plays a central role.

A final limitation of the our approach concerns reduced model interpretability. While the primary goal of the selection model is to obtain

consistent estimates in the outcome equation, it is still desirable to understand the underlying selection mechanism. Unlike traditional Probit models, which offer interpretable coefficients that link predictors to selection probabilities, XGBoost does not provide such direct interpretability. To address this, we rely on Shapley values as variable importance tools, which offer a post hoc decomposition of each predictor's contribution to the model's predictions. This allows us to recover some degree of transparency regarding the factors driving selection. In that respect, we face the well-known trade-off inherent to machine learning models: a loss in interpretability is exchanged for the ability to uncover complex, non-linear relationships—such as the predictive role of missingness—that traditional linear models are generally unable to capture.

3.7 Concluding remarks

This study extends the traditional Tobit 2 model by integrating machine learning techniques to improve the estimation of firm growth dynamics when working with incomplete data and selection bias. Our approach tackles key limitations of traditional Heckman selection models, providing greater flexibility in capturing non-linear relationships and the mechanisms underlying growth observability. By applying this framework to ORBIS data on Italian manufacturing firms, we show that tree-based methods can significantly improve the predictive performance of the selection equation, outperforming traditional probit models. Our findings support a persistent rejection of Gibrat's law, indicating that firm growth is negatively related to size, regardless of the selection mechanism. This result holds across different growth measures, such as revenues and employment, though the strength of the relationship varies. Notably, the positive contribution of innovation becomes evident when machine learning methods are incorporated into the Heckman framework. This suggests that traditional parametric models may underestimate the role of innovation due to their limited capacity to account for selection dynamics and missing data patterns. While our approach provides new insights for handling selection bias and missingness, it is not

without limitations. The use of gradient boosting in the first stage introduces challenges related to the estimation of the IMR, particularly due to complexity of MLE and the restricted support compared to traditional models. Bayesian nonparametric methods offer a flexible alternative for modeling selection mechanisms without relying on strong distributional assumptions. If suitably equipped with tools to handle and incorporate the signal from missing data in the model, they could represent a significant advancement. In conclusion, this paper contributes to the growing literature on the integration of machine learning in econometric models, highlighting both its potential and its limitations. By bridging the gap between traditional selection models and cutting-edge predictive algorithms, we offer new insights into the dynamics of firm growth and provide a perspective for future methodological advancements in the analysis of censored and incomplete data.

Conclusions

This thesis sheds light on the role of statistical learning in analyzing and predicting firm dynamics, stressing the predictive power of machine learning techniques for determining failure risk, success probabilities, and potential growth. A central theme of this thesis is the challenge posed by unobservable firm characteristics that affect future firm dynamics and the methods used to control for them. At the same time, the thesis highlights the strong persistence of observable firm conditions over time, underscoring their importance in shaping long-term firm outcomes. By integrating machine learning with traditional econometric models, this research advances our understanding of firm behavior and provides insights relevant to policy and financial decision-making.

The first chapter shows how machine learning can extract meaningful insights from financial indicators and classify firms into risk categories based on past failures. In particular, extreme gradient boosting outperforms traditional econometric and machine-learning methods, especially in handling missing financial data. The algorithm improves prediction accuracy over conventional credit scoring tools such as Z-scores and DtD. Based on its strongly accurate predictions, we define "zombie firms" as those that persist in the highest risk decile for at least three consecutive years. Identifying these firms is crucial for financial institutions to allocate credit efficiently and mitigate adverse selection problems inherent in imperfect financial markets. More broadly, this work

helps assessing financial distress at both the firm and systemic levels, reinforcing the evidence that zombie firms slow economic growth by misallocating resources. Indeed, our findings indicate that Italian zombie firms are mostly small and show low prod. Their numbers rise during recessions and fall during periods of economic expansion or recovery. While this work does not make direct post-pandemic predictions, it underscores the importance of distinguishing viable firms from non-viable ones, especially as public support programs come to an end.

The second chapter of the dissertation examines the dynamics of firm success, and the role of early-life conditions in shaping later performances. We find that, While early-life conditions do not have a direct causal effect on later economic success when controlling for contemporaneous characteristics, they exhibit strong predictive power due to their correlation with later firm attributes. The findings confirm that firm success probabilities are highly persistent. Firms that perform well early on tend to maintain their advantage, while those struggling in their formative years face an increased risk of exit. This persistence is largely driven by selection effects rather than intra-firm learning. The results suggest that most of the variation in young firms' success rates can be explained by early-life conditions, indicating that external economic factors play a secondary role. Despite these insights, some aspects of firm learning remain unobserved, and further research on the role of innovation, managerial skills, and strategic flexibility is needed. These findings underscore the importance of policies that enhance early firm conditions, such as improved access to finance, mentorship programs, and startup support, as firm trajectories are largely determined in their formative years.

The final chapter extends the discussion on firm dynamics to firm growth, by integrating machine learning techniques into traditional Tobit 2 models. This approach improves the estimation of firm growth determinants by addressing selection bias and missing data more effectively than standard econometric models. Applying this methodology to Italian manufacturing firms, this work further supports that tree-based

methods significantly enhance predictive accuracy, outperforming standard probit models within the Heckman model. A key empirical result is the persistent rejection of Gibrat's law: firm growth is inversely related to size, regardless of the selection mechanism. This relationship holds across different growth horizons, with innovation playing a crucial role—though its impact is often underestimated by traditional models due to their limited capacity to capture selection dynamics. While this contribution provides a more flexible framework for analyzing censored data, it introduces challenges in estimating the IMR with maximum likelihood. Future research could explore Bayesian nonparametric methods as a more adaptable alternative. More broadly, this work contributes to the growing literature on the integration of machine learning into econometrics, highlighting both its potential and its limitations in addressing selection bias and incomplete data.

A common thread across the chapters is that machine learning algorithms not only enhance predictive accuracy and help overcome econometric limitations, but also provide valuable tools for addressing key questions related to economic policy. The findings underscore the need for early-stage interventions to foster firm success, as well as improve resource allocation. Future research can further enhance our understanding of firm behavior and contribute to more effective policies.

Appendix A

Appendix to Chapter 1

A.1 Additional tables

Table A.1.1: Panel (A): List of predictors for firms' failures.

Variables	Description
Material Costs (FA), Costs of Employees (FA), Added Value (FA), Taxation, Tax and Pensions' Payables (FA), Revenues (FA), Financial Expenses (FA), Interest Payments (FA), Cash Flow (FA), Fixed Assets (FA), Current Assets (FA), Shareholders' Funds (FA), Retained Earnings (FA), Long-Term Debt (FA), Loans (FA), Current Liabilities (FA), EBITDA (Earnings before interest, Taxation, Depreciation and Amortization) (PFT), Intangible Fixed Assets (FC), Total Assets (SI), Net Income (SI), Number of Employees (SI)	Original accounts (in euros).
Capital Intensity (FA)	Ratio of fixed assets over the number of employees.

Table A.1.1: Panel (B): List of predictors for firms' failures.

Variables	**Description**
Corporate Control (G)	Binary variable equal to one if a firm belongs to a corporate group.
Consolidated Accounts (G)	Binary variable equal to one if the firm consolidates accounts of its subsidiaries.
Number of Patents (I)	Portfolio of patents granted to a firm by patent offices (*Dummy Patents* equal to 0 if the firm issued no patents, and 1 otherwise).
Number of Trademarks (I)	Total number of trademarks issued to the firm by national or international trademark offices (*Dummy Trademarks* equal to 0 if the firm issued no trademarks, and 1 otherwise).
NACE rev. 2 (SE)	4-digit industry affiliation following European classification NACE rev. 2.
NUTS 2 regions (A)	Region in which the company is located.
TFP (PDT)	Total factor productivity computed as in Ackerberg, Caves, and Frazer, 2015.

Table A.1.1: Panel (C): List of predictors for firms' failures.

Variables	**Description**
Interest Benchmarking (ZI)	Proxy for *zombies* proposed by Caballero, Hoshi, and Kashyap (2008) and calculated as $$R^* = rs_{t-1}BS_{i,t-1} + \left(\frac{1}{5}\sum_{j=1}^{5} rl_{t-j}\right) BL_{i,t-1} + rcb_{5y,t} \cdot Bonds_{i,t-1}$$ where $BS_{i,t-1}$ are short-term bank loans, $BL_{i,t-1}$ are long-term bank loans, rs_{t-1} is the average short-term prime rate in year t, rl_{t-j} is the average long-term prime rate in year t, $Bonds$ are the total outstanding bonds, and $rcb_{5y,t}$ is the minimum observed rate on any convertible corporate bond issued over the previous five years.
Interest Coverage Ratio (ZI)	Ratio of EBIT over interest expenses. When it is less than one for three consecutive years and the firm is at least ten years old, then Bank of Korea (2013), Müge Adalet McGowan, Andrews, and Millot (2018), and Banerjee and Hofmann (2018) assume a firm is a *zombie*.

Table A.1.1: Panel (D): List of predictors for firms' failures.

Variables	**Description**
Financial Misallocation (ZI)	Binary indicator adopted by Schivardi, Sette, and Tabellini, 2021 for capturing *zombie lending*, based on both $ROA = \frac{\frac{1}{3}\sum_{t=1}^{3} EBITDA_t}{Total\ Assets} < prime$ and $Leverage = \frac{Financial\ Debt}{Total\ Assets} > \tilde{L}$ where $prime$ is the measure of the cost of capital for firms with a Z-score equal to 1 or 2, and where $\tilde{L}$ is the median value of leverage in the current year for firms that exited in the following two years.
Negative Value Added (ZI)	Binary variable to identify *zombie firms* (Müge Adalet McGowan, Andrews, and Millot, 2018), equal to one when the value added is negative, i.e., when the value of sold output is less than purchases of intermediate inputs.
Profitability (ZI)	Ratio of EBITDA over total assets, adopted by Schivardi, Sette, and Tabellini (2021) as a control for *zombie lending*.
Financial Constraints (FC)	Proxy of financial constraints as in Nickell and Nicolitsas, 1999, calculated as the ratio between interest payments and cash flow.

Table A.1.1: Panel (E): List of predictors for firms' failures.

Variables	**Description**
Size-Age (FC)	Synthetic indicator proposed by Hadlock and Pierce, 2010: $-0.737 \log(Total\ Assets) + 0.043 \log(Total\ Assets)^2 - 0.040 Age$
Financial Sustainability (FC)	It is a ratio calculated as Financial Expenses over Operating Revenues.
Capital Adequacy Ratio (FC)	Ratio of shareholders' funds over short and log term debt.
Liquidity Ratio (FA)	Ratio of current assets (net of stocks) over current liabilities.
Solvency Ratio (FA)	Ratio of shareholders' funds over current and non-current liabilities.
Liquidity Returns (FA)	Ratio of cash flow over total assets.
Tax and Pension Payables (FA)	Ratio of the sum of tax and pension payables over total assets.

Table A.1.2: 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)	
Financial Constraint : 0	37% (105904)	49% (8419)	χ^2_1=968.27, P<0.001
Financial Constraint : 1	63% (181683)	51% (8900)	
Financial Misallocation : 0	39% (112560)	54% (9276)	χ^2_1=1415.82, P<0.001
Financial Misallocation : 1	61% (175027)	46% (8043)	
Total Factor Productivity : 0	36% (104345)	38% (6600)	χ^2_1=23.52, P<0.001
Total Factor Productivity : 1	64% (183242)	62% (10719)	
Solvency Ratio : 0	41% (118851)	63% (10897)	χ^2_1=3115.5, P<0.001
Solvency Ratio : 1	59% (168736)	37% (6422)	
Liquidity Ratio : 0	42% (119357)	72% (12543)	χ^2_1=6362.72, P<0.001
Liquidity Ratio : 1	58% (168230)	28% (4776)	
Size-Age : 0	42% (120260)	75% (12989)	χ^2_1=7310.19, P<0.001
Size-Age : 1	58% (167327)	25% (4330)	
Liquidity Returns : 0	39% (112561)	54% (9277)	χ^2_1=1416.88, P<0.001
Liquidity Returns : 1	61% (175026)	46% (8042)	
Labour Productivity : 0	34% (97253)	36% (6221)	χ^2_1=32.23, P<0.001
Labour Productivity : 1	66% (190334)	64% (11098)	
Profitability : 0	37% (105907)	49% (8422)	χ^2_1=970.93, P<0.001
Profitability : 1	63% (181680)	51% (8897)	
Financial Sustainability : 0	41% (117294)	64% (11119)	χ^2_1=3673.94, P<0.001
Financial Sustainability : 1	59% (170293)	36% (6200)	
Capital Intensity : 0	36% (104122)	42% (7325)	χ^2_1=261.17, P<0.001
Capital Intensity : 1	64% (183465)	58% (9994)	

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

Table A.1.3: *Zombies*, geography and selected financial indicators

		Interest payments			Value generation		
Region	NUTS 2	Common support	Zombies	ICR<1	Common support	Zombies	Value added<0
Abruzzo	ITF1	0.16	0.48	0.36	0.15	0.57	0.27
Basilicata	ITF5	0.19	0.53	0.29	0.12	0.66	0.22
Calabria	ITF6	0.20	0.50	0.30	0.13	0.64	0.23
Campania	ITF3	0.16	0.57	0.27	0.12	0.62	0.25
Emilia-Romagna	ITH5	0.14	0.29	0.57	0.14	0.46	0.39
Friuli-Venezia Giulia	ITH4	0.15	0.26	0.59	0.14	0.49	0.37
Lazio	ITI4	0.15	0.55	0.31	0.13	0.62	0.25
Liguria	ITC3	0.20	0.37	0.43	0.14	0.54	0.32
Lombardia	ITC4	0.15	0.28	0.57	0.14	0.48	0.38
Marche	ITI3	0.17	0.41	0.42	0.16	0.53	0.31
Molise	ITF2	0.13	0.46	0.41	0.15	0.57	0.29
Piemonte	ITC1	0.15	0.29	0.56	0.15	0.46	0.39
Puglia	ITF4	0.19	0.49	0.33	0.15	0.58	0.26
Sardegna	ITG2	0.17	0.30	0.52	0.15	0.54	0.31
Sicilia	ITG1	0.18	0.41	0.41	0.14	0.53	0.33
Trentino-Alto Adige	ITH1/2	0.18	0.46	0.36	0.17	0.53	0.30
Toscana	ITI1	0.16	0.25	0.59	0.12	0.42	0.46
Umbria	ITI2	0.21	0.36	0.43	0.17	0.46	0.37
Valle d'Aosta	ITC2	0.06	0.28	0.66	0.14	0.45	0.41
Veneto	ITH3	0.15	0.30	0.54	0.15	0.45	0.40
Num. obs.			30,380			24,351	

Notes: We report the proportions of *zombies* against firms with ICR lower than 1 and against firms with negative value added for each NUTS 2-digit region in Italy. Common support indicates the overlap between *zombie firms* and the firms that have problems with interest payments ($ICR < 1$) and have a negative value added, respectively.

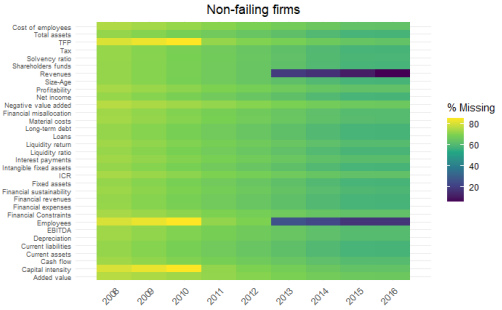
Table A.1.4: *Zombies*, industries and viability indicators

		Interest payments			Value generation		
Manufacturing of	NACE	Common support	Zombies	ICR< 1	Common support	Zombies	Value added< 0
Food products	10	0.18	0.31	0.50	0.15	0.45	0.40
Beverages	11	0.33	0.33	0.33	0.12	0.50	0.38
Tobacco products	12	0.19	0.35	0.46	0.16	0.55	0.29
Textiles	13	0.19	0.49	0.31	0.17	0.51	0.32
Wearing apparel	14	0.15	0.53	0.32	0.13	0.55	0.32
Leather	15	0.17	0.34	0.49	0.17	0.49	0.34
Wood and cork products	16	0.12	0.31	0.57	0.13	0.56	0.31
Paper products	17	0.15	0.32	0.53	0.16	0.54	0.30
Printing and reproduction of recorded media	18	0.14	0.34	0.52	0.11	0.44	0.45
Coke and refined petroleum products	19	0.13	0.34	0.53	0.14	0.46	0.41
Chemicals and chemical products	20	0.19	0.36	0.45	0.10	0.49	0.41
Pharmaceutical products and preparations	21	0.12	0.30	0.57	0.13	0.53	0.34
Rubber and plastic products	22	0.18	0.31	0.52	0.17	0.51	0.32
Other non metallic mineral products	23	0.12	0.26	0.61	0.12	0.49	0.39
Basic metals	24	0.14	0.39	0.46	0.14	0.59	0.27
Fabricated metal products	25	0.16	0.41	0.44	0.11	0.53	0.35
Computer, electronic and optical products	26	0.16	0.39	0.44	0.14	0.54	0.32
Electrical equipment	27	0.15	0.37	0.48	0.14	0.52	0.34
Machinery and equipment n.e.c.	28	0.17	0.33	0.50	0.16	0.50	0.35
Motor vehicle, trailers and semi-trailers	29	0.18	0.45	0.37	0.18	0.49	0.32
Other transport equipment	30	0.19	0.35	0.46	0.19	0.52	0.30
Forniture	31	0.19	0.40	0.41	0.16	0.51	0.33
Other manufacturing	32	0.12	0.51	0.37	0.10	0.55	0.35
Num. obs.		30,380			24,351		

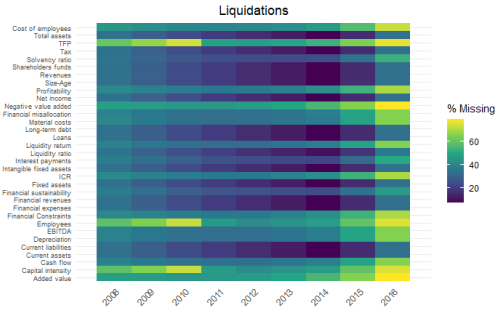
Notes: We report for 2-digit industries (NACE Rev. 2) the shares of zombies compared to firms with an ICR of less than 1 and compared to firms with negative value added. The common support column contains the proportions of firms classified as *zombies* by our algorithm that have either an ICR of less than 1 or negative value added.

A.2 Additional figures

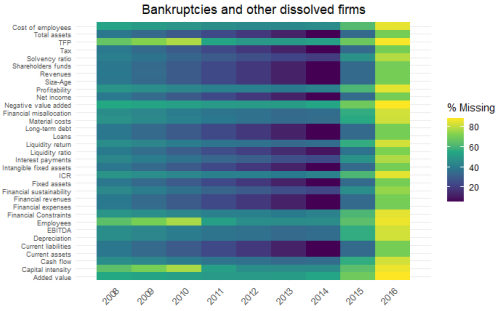
Figure A.2.1: Patterns of missingness over time across firms



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

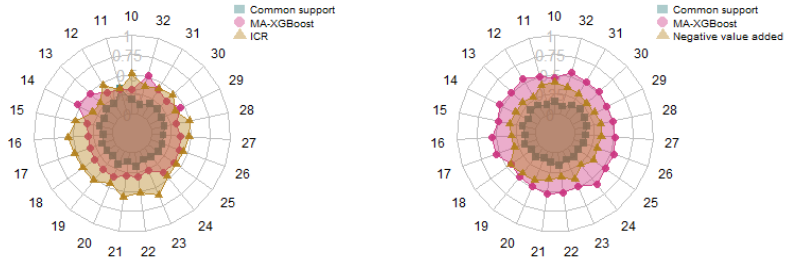


Panel (B): Share of missing values out of 7,221 firms in liquidation.



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

Figure A.2.2: *Zombie firms* and industries



(a) *Zombies* and interest payments

(b) *Zombies* and value destruction

Notes: The rays of the radar show, at the industry level (NACE Rev. 2, 2-digit manufacturing codes), *zombie firms* against firms that report ICR of less than 1 (panel a) and the proportion of firms reporting 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 with XGboost. The triangles indicate the fraction of firms identified with ICR ¡ 1 or negative value added, alternatively. Along each ray, the values of the squares, circles, and triangles sum to 1. Panels (a) and (b) include a total of 30,380 and 24,351 observations, respectively. See Table A.1.4 for the legend of the NACE Rev. 2, 2-digit manufacturing codes.

A.3 Methodologies

A.3.1 XGBoost

XGBoost is a gradient-boosting algorithm introduced by T. Chen and Guestrin, 2016. 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:

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

where $\mathcal{T}_j \left(\mathbf{X}_{i,t-1}; \mathcal{D}_j, \mathcal{W}_j \right)$ 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.

XGBoost is characterized by the fact that it includes several tools that increase the training speed and deal with sparsity in the data (Bentéjac, Csörgő, and Martínez-Muñoz, 2021). In traditional algorithms for splitting data, finding the best splits is quite time consuming. To be effi-

cient, each variable is sorted and all potential splits are explored to determine the optimal one. XGBoost overcomes this problem by using a compressed column-based structure with parallel processing to avoid repeated sorting of the data. The algorithm for finding the best split is now based on the percentiles of the features, so only a subset of the possible splits is examined, which significantly increases the training speed. In addition, XGBoost uses *default directions* to handle sparsity patterns. For missing values, if a feature required for a split is missing, XGBoost assigns the observation to a default direction (left or right) learned from the data. This approach is useful for handling and incorporating information about missing values into the model.

A.3.2 BART-MIA

BART is a Bayesian sum-of-trees ensemble algorithm and its approach is based on a fully Bayesian probability model (Kapelner and Bleich, 2015; Kapelner and Bleich, 2016). The revised version of the BART model we use for our predictions can be expressed as follows:

$$Y_{i,t} = \sum_{j=1}^{J} \mathcal{T}_j(\mathbf{X}_{i,t-1}; \mathcal{D}_j, \mathcal{M}_j) + \epsilon_{i,t-1}, \qquad \epsilon_{i,t-1} \sim \mathcal{N}(0, \sigma^2), \qquad \text{(A.1)}$$

where each $\mathcal{T}_j(\mathbf{X}_{i,t-1}; \mathcal{D}_j, \mathcal{M}_j)$ denotes a unique binary tree. Each $\mathcal{T}_j$ is a function that sorts each unit into one of the sets of m_j terminal nodes associated with mean parameters $\mathcal{M}_j\{\mu_1, ..., \mu_{m_j}\}$ based on a set of decision rules, $\mathcal{D}_j$. The error terms $\epsilon_{i,t-1}$ are usually assumed to be independent and identically normally distributed if the outcome is continuous (Chipman, George, McCulloch, et al., 2010). The Bayesian component of the algorithm is incorporated in a set of three different priors on: (i) the structure of the trees, $\mathcal{D}_j$ (this prior aims to limit the complexity of each tree $\mathcal{T}$ and serves as a regularization tool); (ii) the distribution of the outcome in the nodes, $\mathcal{M}_j$ (this prior aims to reduce the node predictions to the center of the distribution of the response variable Y); (iii) the error

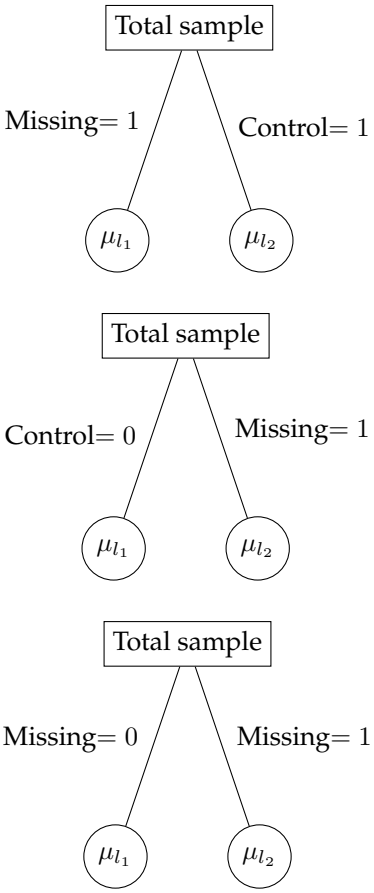
variance σ^2 (which bounds away σ^2 from very small values that would cause the algorithm to overfit the training data)[1]. The goal of these priors is to regularize the algorithm and prevent individual trees from dominating the overall fit of the model. This property is considered important to balance the tendency of tree-based methods to overfit the training data (Kapelner and Bleich, 2016).

BART-MIA extends the original BART algorithm by including additional information from patterns of missing values (Kapelner and Bleich, 2015). This is done by introducing the possibility to share on a feature for missing values in each binary tree component of the BART algorithm, $\mathcal{T}$. Figure A.3.1 illustrates the intuition behind the MIA procedure in a simple case with only one variable (*Control*). In the presence of missing values in this variable, the algorithm creates a new feature (*Missing*) – namely, an indicator variable that takes the value 1 when the ith observation for the variable is missing – and uses it to perform the data split. As shown in A.3.1, the algorithm in this case has three possible splits. As in the usual CART procedure, the algorithm chooses the one that minimizes the prediction error in the generated leaves l_1 and l_2.

As shown by Twala, Jones, and Hand (2008), this splitting rule allows trees to better capture the direct influence of missing values as another predictor of the response variable. Furthermore, after a theoretical and empirical comparison of different missing value strategies in trees, Josse et al., 2019 show that MIA can handle both non-informative and informative missing values.

[1]The choice of priors and the derivation of posterior distributions are discussed in detail in Chipman, George, McCulloch, et al., 2010 and Kapelner and Bleich, 2016. Namely: (i) the prior on the probability that a node splits at depth k is $\beta(1k)^{-\eta}$, where $\beta \in (0,1), \eta \in [0,\infty)$ (these hyperparameters are generally chosen to be $\eta 2$ and $\beta 0.95$); (ii) the prior on the probability distribution in the nodes is a normal distribution with zero mean: $\mathcal{N}(0,\sigma_q^2)$ where $\sigma_q\sigma_0/\sqrt{q}$ and σ_0 can be used to calibrate the plausible range of the regression function; (iii) the prior on the error variance is $\sigma^2 \sim InvGamma(v/2, v\lambda/2)$ where λ is determined from the data such that the BART improves the RMSE of an OLS model in 90% of cases.

Figure A.3.1: MIA splitting rules



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

A.4 Selection of predictors with LASSO

In our analysis, we asserted that we need as much in-sample information as possible to reduce out-of-sample prediction errors. However, especially in small sample analyses, overfitting problems can arise as the dimensionality of the inputs increases. The so-called *curse of dimensionality* is an obstacle when working with finite data samples and many variables. The basic reference is the work of Bellman, 1961, who introduced the notion of dimensionality reduction. Falco Joannes Bargagli-Stoffi, Cevolani, and Gnecco, 2020 and Falco J. Bargagli-Stoffi, Cevolani, and Gnecco, 2022 discuss the role of regularization and dimensionality reduction in the context of machine learning. In this Appendix, we show what happens when we sift through firm-level data and attempt to extract a set of predictors with the highest ability to detect financial distress. The natural candidate for reducing the dimensionality of a matrix of predictors is the LOGIT-LASSO (Ahrens, Christian B Hansen, and Mark E Schaffer, 2019), whose functional form in a panel setting is the following:

$$\begin{aligned} \underset{\beta \in \mathbb{R}^p}{\arg\min} \quad & \frac{1}{2N} \sum_{i=1}^{N} \left(y_{i,t}(x_{i,t-1}^T \beta) - \log(1 + e^{(x_{i,t-1}^T \beta)}) \right)^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq k. \end{aligned} \tag{A.2}$$

where $y_{i,t}$ is a binary variable equal to one if a firm i failed at time t and zero otherwise. Each $x_{i,t-1}$ is a lagged predictor chosen in $\mathbb{R}^p$ at time $t-1$, while $\|\beta\|_1 = \sum_{j1}^{p} |\beta_j|$. The constraint $\|\beta\|_1 \leq k$ limits the complexity of the model to avoid overfitting, and k is chosen following Ahrens, Christian B Hansen, and Mark E Schaffer (2019) as the value that maximizes the Extended Bayesian Information Criteria (J. Chen and Z. Chen, 2008). We use the rigorous penalization introduced by Belloni, Chernozhukov, C. Hansen, et al., 2016 to account for the possible presence of heteroskedastic, non-Gaussian, and cluster-dependent errors. Table A.4.1 shows the ten highest ranked predictors.

Although the number of predictors varies over time, up to a maximum of twenty-one characteristics in 2015, we find a core set of frequently selected predictors. This stable set includes indicators of financial distress (*Liquidity Returns, ICR, Interest Benchmark, Financial Constraint*) and indicators of firms' core economic activities (*Negative value added, TFP, Size-Age*). Apparently, firms controlled by parent companies (*Corporate Control*) are less likely to fail in each period, i.e. the predictor enters the algorithm with a negative coefficient[2]. The same is true for the *Dummy Trademarks* and the *Dummy Patents* since it makes sense that intangible assets reduce the probability of exit.

Some of the top predictors we showed were used to measure either financial distress or *zombie lending*. However, the rankings change over time, and we cannot discern a meaningful pattern in these changes. In this context, we cannot rely on a single indicator (or set of them) to derive predictions about defaults. If we did, we would have a higher rate of *false positives* (when our forecasts suggest that a firm is at risk of failure, but it is not) and a higher rate of *false negatives* (when forecasts incorrectly suggest that a firm in trouble deserves a loan).

Indeed, at this stage, we cannot rule out the possibility that a different ranking over time is due to better use of the newly acquired information from the in-sample information after new failures have been observed. It could also be that a change in rankings reflects a change in the business environment in which firms operate. We presume applications to different aggregates (countries, regions, industries) may result in different rankings. We conclude that it is better to keep the full battery of predictors, provided there is no dimensionality problem in the predictors when we can train with a large number of observed firm-level outcomes.

[2]For an in-depth analysis of the impact of corporate control on the performance of Italian firms, see Riccaboni, X. Wang, and Z. Zhu (2021).

Table A.4.1: Top 10 predictors for firms' failures - Results from a rigorous LOGIT-LASSO

Rank	2017	2016	2015	2014	2013	2012	2011	2010	2009
1	Liquidity Returns	Negative value added	Negative value added	Negative value added	Liquidity Returns	Negative value-added	Negative value added	Negative value added	Negative value added
2	Negative value added	Financial Constraint	Corporate Control	Liquidity Returns	Negative value added	Profitability	Liquidity Returns	Liquidity Returns	Liquidity Returns
3	Corporate Control	ICR	Financial Constraint	Profitability	Profitability	Financial Constraint	Financial Constraint	Profitability	Financial Constraint
4	Financial constraint	Corporate Control	ICR	Solvency Ratio	Solvency Ratio	Corporate Control	Corporate Control	Financial Constraint	Profitability
5	ICR	Solvency Ratio	Profitability	Financial Constraint	Corporate Control	Solvency Ratio	Solvency Ratio	Corporate Control	Corporate Control
6	Profitability	Size-Age	Solvency Ratio	Corporate Control	Financial Constraint	ICR	Size-age	Solvency Ratio	Solvency Ratio
7	Solvency Ratio	Liquidity Ratio	Size-age	ICR	Size-Age	Liquidity Returns	ICR	Area	ICR
8	Size-age	Liquidity Returns	Liquidity Ratio	Size-age	ICR	Size-age	Misallocated fixed	Dummy Trademarks	Dummy Trademarks
9	Liquidity Ratio	Capital Intensity	Area	TFP	Dummy Patents	Dummy Patents	Dummy Trademarks	ICR	Size-age
10	Capital Intensity	TFP	Capital Intensity	Liquidity Ratio	TFP	Dummy Trademarks	Dummy Patents	Dummy Patents	Capital Intensity

Rankings are determined after conducting a rigorous LOGIT-LASSO (Ahrens, Christian B. Hansen, and Mark E. Schaffer, 2020; Belloni, Chernozhukov, and Wei, 2016) each year on the entire battery of predictors described in Figure A.1.1. Only the first ten selections are reported. The procedure selects a different number of predictors each year, up to a maximum of 21.

A.5 Imputation of missing predictors

In this appendix, we show the performance of machine learning algorithms when we impute the missing values of predictors using two possibilities known in the literature: out-of-range and median imputation. In this way, we can use all the observations available in the dataset. We first conduct the analysis keeping all categories of firm failures and then exclude the cases of liquidations since the latter can sometimes have reasons other than financial distress. The training and test datasets include 238,148 and 59,537 observations, respectively, in each iteration when liquidations are excluded. On the other hand, the training and testing datasets include 243,924 and 60,982 observations, respectively, when liquidations are included. In Table A.5.1 we perform out-of-range imputation by imputing all missing values in the dataset with an out-of-range value, i.e., 1020. In Table A.5.2 we perform median imputation by imputing the missing values of a variable q with its median $q_{.5}$.

Table A.5.1: Performance of the model with out-of-range imputed data

(a) Liquidations excluded

Method	ROC	PR	F1-Score	BACC	R^2	Time
Logit	0.9657	0.6689	0.1270	0.7588	0.4558	56.49
Ctree	0.9712	0.6999	0.1323	0.7686	0.4840	1628.08
Random Forest	0.9343	0.7000	0.2011	0.8371	0.4999	608.33
XGBoost	0.9785	0.7592	0.1270	0.7586	0.5427	40.28
BART	0.9770	0.7482	0.1296	0.7635	0.5327	2736.74
Super Learner	0.9788	0.7574	0.1289	0.7626	0.5454	11765.18

(b) Liquidations included

Method	AUC	PR	F1-Score	BACC	R^2	Time
Logit	0.9496	0.6482	0.2039	0.7647	0.4173	76.89
Ctree	0.9652	0.7230	0.2264	0.7919	0.48907	2835.23
Random Forest	0.9410	0.7313	0.2680	0.8214	0.5130	695.48
XGBoost	0.9733	0.7815	0.2039	0.7648	0.5525	45.29
BART	0.9719	0.7718	0.2048	0.7659	0.5412	4056.36
Super Learner	0.9748	0.7896	0.2039	0.7649	0.5620	13992.57

Notes: All algorithms were trained with five-fold cross-validation. All metrics correspond to the five-fold average. Time indicates the average seconds taken to train the model in each fold.

Table A.5.2: Performance of models with column-wise median imputed data

(a) Liquidations excluded

Method	ROC	PR	F1-Score	BACC	R^2	Time
Logit	0.8825	0.2690	0.1465	0.7540	0.1495	50.77
Ctree	0.9270	0.4017	0.1492	0.7962	0.2364	1975.38
Random Forest	0.7767	0.4337	0.1487	0.7060	0.2667	855.37
XGBoost	0.9584	0.5246	0.1271	0.7579	0.3281	39.51
BART	0.9485	0.4936	0.1328	0.7688	0.3014	4345.19
Super Learner	0.9595	0.5393	0.1269	0.7576	0.3389	13927.68

(b) Liquidations included

Method	ROC	PR	F1-Score	BACC	R^2	Time
Logit	0.8907	0.3966	0.2282	0.7667	0.2302	60.44
Ctree	0.9237	0.5059	0.2187	0.7808	0.3082	2429.78
Random Forest	0.8281	0.5453	0.2227	0.7304	0.3420	1167.71
XGBoost	0.9536	0.6175	0.2037	0.7636	0.3942	40.93
BART	0.9440	0.5890	0.2063	0.7663	0.3696	4702.54
Super Learner	0.9551	0.6295	0.2043	0.7643	0.4051	15218.50

Notes: All algorithms are trained with five-fold cross-validation. All metrics correspond to the five-fold average. The time indicates the average seconds taken to train the model in each fold.

Appendix B

Appendix to Chapter 2

B.0.1 Predictive performance compared to OLS

Table B.0.1 compares our primary model, 2S-XGB, with two traditional econometric benchmark models: Ordinary Least Squares (OLS) and logistic regression (Logit). Since both OLS and Logit cannot handle missing values, the subsequent horse-race analysis is conducted using a more limited set of predictors. Specifically, worker and DGA variables were excluded due to a high proportion of missing values caused by incomplete matching of worker-related information across many startups. Additionally, the variable importance analysis in the main text revealed their low predictive power, further justifying their exclusion.

The horse race shows that the 2S-XGB method consistently predicts economic success better than either Logit or OLS, regardless of the age of the firm. The improvement is most visible when the R-squared is used to evaluate predictive performance and when firms are young. For instance, the R-Squared of 2S-XGB is about 52 per cent when economic success at age two is predicted from initial life conditions, an improvement in the predictive performance of about 93 per cent compared to Logit (R-Squared = 0.27) and of about 126 per cent compared to OLS

(R-Squared = 0.23). At age ten, the R-Squared of the 2S-XGB model is 21 percent, an improvement of 17 percent compared to the Logit model (R-Squared = 0.18) and of about 31 percent compared to the OLS model (R-Squared = 0.16).

In terms of the AUC-ROC, all models predict economic success reasonably well to very well, depending on the age of the firm. In particular, the AUC of the 2S-XGB model is higher yet similar to that of the Logit model for all ages. The AUC of the OLS model is always below that of the 2S-XGB and Logit models, indicating that these models best separate firms likely to meet the threshold from firms that do not.

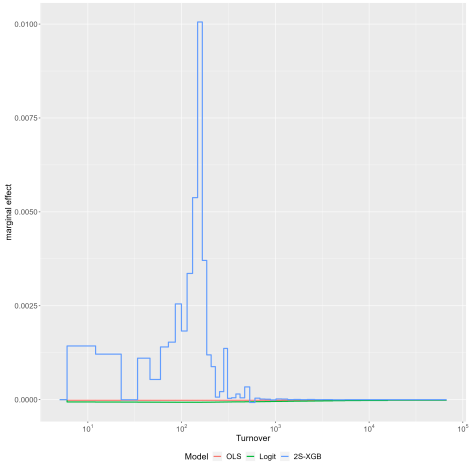
Table B.0.1: Predicting economics success: Models horse-race

Model	Gross train	Net train	Gross test	Net test	AUC-ROC	AUC-PR	R-Squared	Age of success	Age features
2S-XGB	84,919	82,591	198,086	192,848	.8997	.9234	.5234	2	1
Logit	84,919	82,591	198,086	192,848	.8108	.8674	.2730	2	1
OLS	84,919	82,591	198,086	192,848	.7955	.8591	.2282	2	1
2S-XGB	84,919	33,638	198,086	78,337	.8080	.9084	.2794	6	1
Logit	84,919	33,638	198,086	78,337	.7711	.8978	.1868	6	1
OLS	84,919	33,638	198,086	78,337	.7594	.8927	.1603	6	1
2S-XGB	84,919	15,073	198,086	35,075	.7796	.9164	.2146	10	1
Logit	84,919	15,073	198,086	35,075	.7730	.9180	.1807	10	1
OLS	84,919	15,073	198,086	35,075	.7593	.9124	.1561	10	1

One reason to use 2S-XGB is its ability to exploit interactions and non-linearities. To see how these models differ in the way they use features, we plot the marginal effect of turnover on the probability of being economically successful at age two (Figure B.0.1). It shows that the marginal effect of turnover for Logit and OLS is constant and near zero. Yet for Logit, we see that the marginal effect of turnover on the prediction is not constant, but depends on the value of the feature. Consequently, the marginal effect of a 1000 euro increase in turnover in the 2S-XGB model is a concave step function with a peak around $10^{2.2} \approx 150$, which is close to the turnover cutoff used to define economic success. The model assigns a high marginal effect around this cutoff, but when turnover is sufficiently below or above this threshold, marginal effects are smaller. As all models are trained on the same set of observations, this illustrates the ability of

2S-XGB to exploit non-linearities.

Figure B.0.1: Models comparison: Marginal effect of turnover



B.1 Interpretable machine learning methods

We use Shapley values and ALE plots to interpret our models. Here we describe the details of these methods.

B.1.1 Shapley values

As machine learning models become more sophisticated and opaque—often referred to as ”black boxes”—it is crucial to ensure that these models are not only accurate but also interpretable and transparent. Shapley values, originally rooted in cooperative game theory, offer a powerful and fair framework for attributing a model’s output to its individual features. These values account for the marginal contribution of each feature across all possible combinations, ensuring that the importance of a feature is evaluated within the context of all others. This makes Shapley values particularly effective for explaining the predictions of complex models,

as they satisfy key properties such as efficiency, symmetry, and additivity (Rozemberczki et al., 2022). We estimate Shapley values for the testing firms using the 'fastshap' package, which efficiently approximates these values. The library employs sampling techniques and algorithmic optimizations to evaluate only a subset of all possible combinations of variables, making it feasible to apply Shapley values to large datasets and complex models without prohibitive computational costs. This approach enables us to harness the theoretical strengths of Shapley values while overcoming computational challenges.

B.1.2 ALE plots

We use Accumulated Local Effects plots (ALE plots, see Apley and J. Zhu, 2020) to visualize the local impact of a change in a specific variable on the predicted value. ALE plots segment the support of x into numerous narrow and non-overlapping intervals. The derivative of the dependent variable y with respect to x within each interval is estimated by the mean of the difference in predicted values when x takes on the value at the start and at the end of the interval. This approach of using the predicted values' differences avoids the necessity of directly computing the partial derivative, making it applicable to scenarios where the partial derivative is not defined. The mean differences (per interval) are then aggregated to derive the ALE, representing the change in y when x transitions from its minimum to the maximum value. To center the ALE of x, the average ALE across all x intervals is subtracted. We focus on the local effects derived from the ALE plots. In particular, we apply the following steps:

1. Make $k = 1, \ldots, K$ non-overlapping intervals of variable x, where the length of each interval equals $(x_{max} - x_{min})/K$. We use the default value $K = 40$.

2. Compute the local effects: For each observation i in each interval k:

(a) Replace x_l by the lower bound of the interval and predict $\hat{y}_i^{k,l}$.

(b) Replace x_i by the upper bound of the interval and predict $\hat{y}_i^{k,u}$.

(c) Compute the average difference in prediction in each interval

$$\Delta\bar{\hat{y}}^k = \frac{1}{N^k} \sum_{i \in I_k} \left(\hat{y}_i^{k,u} - \hat{y}_i^{k,l} \right)$$

where N^k is the number of observations in interval k.

3. For Local Effects (LE) plots:

   (a) Plot $\Delta\bar{\hat{y}}^k$ against interval k.

4. For ALE plots:

   (a) Rescale $\Delta\bar{\hat{y}}^k$ by subtracting the average difference over all intervals $\Delta\bar{\hat{y}}^{k^*} = \Delta\bar{\hat{y}}^k - \Delta\bar{\hat{y}}$.

   (b) Plot $\Delta\bar{\hat{y}}^{k^*}$ against k.

Consequently, the LE plot for a fitted linear model will have a constant slope, while a variable that is not selected by a LASSO estimator will have a zero slope. The ALE plot for a feature from a linear model and the ALE plot for a feature that was not selected by a LASSO estimator will have a constant slope equal to zero.

B.2 Additional tables

B.2.1 Descriptives

Table B.2.1: Entry dynamics

Year	New startups	Incumbents	Entry rate
2006	19,507	133,350	14.63
2007	20,277	146,291	13.86
2008	24,093	154,483	15.60
2009	17,139	168,130	10.19
2010	19,329	166,349	11.62
2011	18,623	175,431	10.62
2012	17,539	179,586	9.77
2013	17,670	181,562	9.73
2014	18,523	187,210	9.89
2015	21,185	191,287	11.07
2016	19,193	199,804	9.61
2017	19,489	205,640	9.48
2018	19,378	211,748	9.15
2019	21,392	217,156	9.85
2020	10,149	230,616	4.40

Table B.2.2: Startups' workforce in the first year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Age of the workforce	40.24	39.08	10.57	19.16	70.56	Age in years
Wages	-256.2	-86.4	626.93	-5187.77	0	Thousand €
Hours worked	1346.75	1373.8	626.65	57.12	2246.14	Count
Number of employees	12.46	2	91.32	1	461.27	Count
DGAs' age	44.98	44	10.27	22.34	73.39	Age in years
DGAs' assets	51815.13	581.43	2199250.02	-34.15	5078404.6	Thousand €

Notes: summary statistics of workforce-related variables for startups during their first year. The features are based on all employment relationships within the firm according to SPolis data.

Table B.2.3: Startups' workforce in the fifth year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Age of the workforce	42.03	41	10.4	20.09	71.45	Age in years
Wages	-397.63	-137.32	823.21	-6241.91	0	Thousand €
Hours worked	1411.82	1442.62	566.59	86.24	2301.94	Count
Number of employees	18.67	3	148.56	1	757.49	Count
DGAs' age	47.89	47	9.39	26.15	74.23	Age in years
DGA's assets	17148.77	812.24	1241689.75	-18.24	1591262.26	Thousand €

Notes: summary statistics of workforce-related variables for startups that successfully survived for 5 consecutive years. The features are based on all employment relationships within the firm according to SPolis data.

Table B.2.4: All firms characteristics

(a) Firms in their first year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Total balance sheet	1344.98	404.74	2893.56	6.61	21013.17	Thousand €
Intangible assets	15.33	0	66.61	0	491.09	Thousand €
Tangible assets	305.22	21.24	920.32	0	6930.16	Thousand €
Liquid assets	199.19	39.92	495.48	-1.38	3771.14	Thousand €
Equity	453.4	87.44	1389.61	-1527.45	10269.78	Thousand €
Long term debt	293.99	0	905.24	0	6699.2	Thousand €
Short term debt	298.81	72.21	745.16	0	5694.48	Thousand €
Turnover	2755.92	503.91	7288.73	8.28	56669.56	Thousand €
Result before taxes	108.56	23.11	380.29	-730.81	2722.23	Thousand €
Taxes	-24.21	-1.36	69.19	-534.01	19	Thousand €
Value added	491.11	162.83	1042.29	-82.64	7760.9	Thousand €
EBITDA	-10.52	0	278.07	-1421.08	1413.23	Thousand €
Dividends	23.04	0	114.79	0	955.17	Thousand €
Labor productivity	65.43	35.5	124.94	-29.65	950.46	Thousand €
ROA	0.02	0.05	0.4	-1.95	0.9	Ratio
Solvency ratio	0.53	0.06	2.42	-6.92	15.24	Ratio
Debt ratio	9.65	2.36	26.05	0.02	192.94	Ratio
Age	1.51	1.5	0.23	1.08	2	Number of years
Entry year	2010.28	2009	5.05	2005	2020	Year
Year	2011.28	2010	5.05	2006	2021	Year

(b) Firms in their fifth year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Total balance sheet	1854.51	693.62	3379.34	10.06	21647.09	Thousand €
Intangible assets	16.36	0	71.96	0	564.27	Thousand €
Tangible assets	431.66	41.95	1100.19	0	7408.13	Thousand €
Liquid assets	290.25	76.49	606.5	-0.18	4073.45	Thousand €
Equity	732.65	225.13	1698.32	-1597.41	10818.79	Thousand €
Long term debt	377.57	0	1042.2	0	7045.67	Thousand €
Short term debt	374.95	92.7	859.33	0.08	5947.16	Thousand €
Turnover	3724.29	829.47	8550.2	9.88	58478.9	Thousand €
Result before taxes	137.47	34.48	432.62	-815.01	2696.6	Thousand €
Taxes	-24.13	-2.65	71.28	-468.58	143.49	Thousand €
Value added	683.16	247.62	1268.61	-79.31	8327.71	Thousand €
EBITDA	-4.55	0	329.54	-1558.78	1446.66	Thousand €
Dividends	39.86	0	153.28	0	1130.41	Thousand €
Labor productivity	59.94	40.43	79.12	-29.09	565.65	Thousand €
ROA	0.03	0.05	0.31	-1.79	0.87	Ratio
Solvency ratio	0.61	0.06	2.61	-7.25	15.43	Ratio
Debt ratio	11.97	3.04	28.13	0.03	200.4	Ratio
Age	5.51	5.5	0.22	5.08	6	Number of years
Entry year	2008.51	2007	3.82	2005	2016	Year
Year	2013.51	2012	3.82	2010	2021	Year

Notes: Table B.2.4a reports summary statistics for all 416,455 firms available in the sample during their first year of activity. Additionally, Table B.2.4b provides summary statistics for all firms that successfully survived for 5 consecutive years, comprising 240,524 out of the initial 416,455.

Table B.2.5: Firms' workforce in the first year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Age of the workforce	40.63	39.5	10.34	19.61	70.6	Age in years
Wages	-346.2	-113.62	744.02	-5610.65	0	Thousand €
Hours worked	1384.21	1420.76	601.48	67.15	2473.85	Count
Number of employees	13.32	3	83.7	1	423.99	Count
DGAs' age	46.21	46	10.08	23.18	73.5	Age in years
DGAs' assets	82667.92	646.22	2804835.1	-27.1	8156989.97	Thousand €

Notes: summary statistics of workforce-related variables for all firms during their first year. The features are based on all employment relationships within the firm according to SPolis data.

Table B.2.6: Firms' workforce in their fifth year

Variable	Mean	Median	SD	Mean 1st percentile	Mean 99th percentile	Unit
Age of the workforce	42.27	41.26	10.19	20.31	71.39	Age in years
Wages	-493.72	-169.45	932.39	-6207.79	0	Thousand €
Hours worked	1420.82	1458.51	542.62	99.7	2301.68	Count
Number of employees	18.78	4	121.48	1	613.4	Count
DGAs' age	48.83	48	9.36	26.95	74.32	Age in years
DGAs' assets	26003.52	854.79	1523747.62	-14.18	2468869.49	Thousand €

Notes: summary statistics of workforce-related variables for firms that successfully survived for 5 consecutive years. The features are based on all employment relationships within the firm according to SPolis data.

B.3 Sector-specific cutoffs

		α-quantile							
Sector	Num. obs.	0.05	0.10	0.20	0.40	0.60	0.80	0.90	0.95
Public administration	56	120	280	1,090	4,000	10,750	23,810	46,860	62,400
Consultancy and research	68,174	20	40	80	180	340	910	2,110	4,160
Construction	19,654	60	120	300	990	2,330	5,690	11,660	22,080
Culture, sports and recreation	4,728	30	60	150	390	850	1,990	3,870	7,280
Healthcare and welfare services	13,415	80	160	240	340	520	1,410	2,670	4,660
Wholesale and retail trade	50,525	50	110	300	1,150	3,000	8,540	19,000	35,140
Manufacturing	16,837	60	150	400	1,340	3,190	8,660	18,340	33,960
ICT	19,315	20	50	120	270	640	2,040	4,620	9,300
Agriculture, forestry and fishing	5,873	50	100	240	1,060	2,750	6,530	11,650	18,750
Accommodation and food service activities	8,445	90	190	410	970	1,740	3,220	5,250	8,680
Education	3,363	20	50	100	250	590	1,590	3,540	6,190
Other services	1,969	40	90	170	440	940	2,280	4,310	8,050
Electricity, natural gas, steam and cooled air	582	30	60	130	340	990	3,240	6,520	13,290
Rental of movables goods and other business services	13,864	40	80	180	550	1,570	4,610	10,380	19,470
Transport and storage	7,879	60	130	320	1,280	3,110	8,930	19,030	33,160
Extraction and distribution of water	725	60	140	360	1,440	3,440	9,790	22,510	43,310
Mining	145	30	70	160	460	930	4,430	7,840	23,550

Notes: The table provides the α-quantiles of the 2019 turnover distribution for each sector (in thousand euros) on the whole sample, along with the number of observations per sector.

B.4 Features

Table B.4.1: Variables and description

Variables	Description
Total balance sheet, Intangible assets, Tangible assets, Participations, Long term claims, Short term claims, Stocks, Trade receivables, Trade credit, Liquid assets, Equity, Third parties, Shared third parties, Equalization, Provisions, Long term debt, Short term debt, Turnover, Costs revenue, Costs wages, Results from participations, Interest income, Interest payments, Other financial results, Extra benefits, Extra costs, Balance extra costs benefits, Result before taxes, Taxes, Net result, Dividends, Value added, EBITDA.	Original accounts from balance sheet in thousand euros. Participations are represented by three variables: one captures the total participations of the firm, while the other two distinguish between participations involving Dutch firms and those involving non-Dutch firms.
Labor productivity	$\frac{\text{Value added}}{\text{Hours worked}}$
Leverage	$\frac{\text{Long and short term debts}}{\text{Total balance sheet}}$
Quick ratio	$\frac{\text{Liquid assets}}{\text{Short term debt}}$
Debt ratio	$\frac{\text{Total balance sheet}}{\text{Long and short term debts}}$
Debt-to-equity ratio	$\frac{\text{Long and short term debts}}{\text{Equity}}$
ROA	$\frac{\text{Net results}}{\text{Total balance sheet}}$
Solvency ratio	$\frac{\text{Net result}+\text{Depreciation}}{\text{Long and short term debts}}$
Altman Z-Score	$6.56 * \frac{\text{Total balance sheet - long and short term debts}}{\text{Total balance sheet}} + 3.26 * \frac{\text{Equity}}{\text{Total balance sheet}} + 6.72 * \frac{\text{EBITDA}+\text{Depreciation}}{\text{Total balance sheet}} + 1.05 * \frac{\text{Equity}}{\text{Long and short term debts}}$
Financial distress	$1 \text{ if } \frac{\text{EBITDA}+\text{Depreciation}}{\text{Interest payments}} < 1, \ 0 \text{ otherwise}$

Variables	Description
Age, NUTS 3 Region, Entry date index, Exit rate fixed effect	Demographic variables. The exit rate fixed effect is a control variable that accounts for the exit rate by year and NUTS 3 region.
Number of workers, Individual hours worked, Individual hours paid, Full-time days worked Temporary contracts, Gender, Individual wage, Workers' age, Share of workers classified as DGAS, First job type, Workers' migration background, Assets, Savings, Stocks (own firm), Educational level.	Spolis variables based on the **entire workforce** matched for the firm. Each variable represents the mean across the workers of the firm for a specific year. Specifically: - Gender: share of male workers - Workers' migration background: classified as none, parents from abroad, or grandparents from abroad - Assets, savings, and stocks: average values across the matched workers of the firm - Educational level: classified as low, medium, or high.
Number of DGAs, Individual hours worked, Individual hours paid, Full-time days worked, Temporary contracts, Gender, Individual wage, DGAs' Age, Job type, Country of origin, Assets, Savings, Stocks (own firm), Educational level.	Spolis variables are based on only DGAs matched for the firm. Each variable represents the average (mean) across the DGAs of the firm for a specific year. The variables include: - Gender: Share of male DGAs. - Country of origin: Dutch Antilles and Aruba, Morocco, The Netherlands, other Western countries, other non-Western countries. - Assets, savings, and stocks: average values across the matched DGAs. - Educational level: classified as low, medium, or high.

B.4.1 Survival models

Table B.4.2: Early-life survival prediction

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age to survive	Age features
56,697	56,697	28,348	28,348	198,441	198,441	.7612	.9364	.1434	2	1
56,697	24,422	28,348	12,249	198,441	85,366	.5829	.9374	.0056	6	1
56,697	11,416	28,348	5,718	198,441	39,960	.5535	.9495	.0016	10	1

Table B.4.3: Survival prediction models at age 3

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age to survive	Age features
36,875	29,651	18,331	14,799	128,408	103,222	.6630	.9457	.0303	6	3
36,875	11,416	18,331	5,718	128,408	39,960	.5530	.9485	.0024	10	3

B.4.2 Economic success with pooled cutoffs

Table B.4.4: Early-life economic success prediction models (pooled cutoffs)

(a) $\alpha = 0.05$, $q_\alpha = 30,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.8278	.9373	.3234	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.6979	.9350	.1278	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.6731	.9420	.1336	10	1

(a) $\alpha = 0.10$, $q_\alpha = 70,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.8644	.9487	.4244	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.7511	.9198	.1939	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.7355	.9293	.1745	10	1

(a) $\alpha = 0.20$, $q_\alpha = 150,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9017	.9510	.5257	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.8120	.9267	.2906	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.7906	.9253	.2356	10	1

(a) $\alpha = 0.40$, $q_\alpha = 400,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9415	.9575	.6373	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.8845	.9303	.4691	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.8601	.9272	.4025	10	1

(a) $\alpha = 0.60$, $q_\alpha = 1,190,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9643	.9692	.6901	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.9135	.9347	.5223	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.8874	.9276	.4581	10	1

(a) $\alpha = 0.80$, $q_\alpha = 3,780,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9758	.8538	.6843	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.9293	.7710	.5108	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.8970	.7339	.4283	10	1

(a) $\alpha = 0.90$, $q_\alpha = 8,860,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9818	.8191	.6685	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.9403	.7004	.4756	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.9089	.6463	.3939	10	1

(a) $\alpha = 0.95$, $q_\alpha = 17,910,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55,173	28,348	27,543	198,441	193,200	.9832	.7840	.6375	2	1
56,697	22,444	28,348	11,238	198,441	78,472	.9432	.6294	.4342	6	1
56,697	10,032	28,348	5,063	198,441	35,153	.9102	.5366	.3283	10	1

Table B.4.5: Next year economic success prediction models (pooled cutoffs)

(a) $\alpha = 0.05$, $q_\alpha = 30,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.8278	.9373	.3234	2	1
24,661	22444	12,360	11,238	86,134	78,472	.8323	.9648	.3392	6	5
11,490	10032	5,761	5,063	40,268	35,153	.8283	.9718	.3511	10	9

(a) $\alpha = 0.10$, $q_\alpha = 70,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.8644	.9335	.4244	2	1
24,661	22,444	12,360	11,238	86,134	78,472	.8728	.9626	.4528	6	5
11,490	10032	5,761	5,063	40,268	35,153	.8792	0.9719	.4847	10	9

(a) $\alpha = 0.20$, $q_\alpha = 150,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9017	.9259	.5257	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9144	.9568	.5715	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9147	.9648	.5977	10	9

(a) $\alpha = 0.40$, $q_\alpha = 400,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9415	.9123	.6373	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9571	.9511	.7277	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9595	.9642	.7555	10	9

(a) $\alpha = 0.60$, $q_\alpha = 1,190,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9643	.8943	.6901	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9751	.9427	.7830	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9744	.9535	.8069	10	9

(a) $\alpha = 0.80$, $q_\alpha = 3,780,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9758	.8538	.6843	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9842	.9228	.7828	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9834	.9356	.8132	10	9

(a) $\alpha = 0.90$, $q_\alpha = 8,860,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9818	.8191	.6685	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9869	.9003	.7653	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9844	.8949	.7840	10	9

(a) $\alpha = 0.95$, $q_\alpha = 17,910,000$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	55173	28,348	27,543	198,441	193,200	.9832	.7840	.6375	2	1
24,661	22444	12,360	11,238	86,134	78,472	.9882	.8729	.7399	6	5
11,490	10032	5,761	5,063	40,268	35,153	.9864	.8739	.7549	10	9

B.4.3 Economic success with sector-specific cutoffs

Table B.4.6: Early-life economic success prediction models (sector-specific cutoffs)

(a) $\alpha = 0.05$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.8283	.9353	.3247	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.7023	.9330	.1320	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.6872	.9447	.1351	10	1

(a) $\alpha = 0.10$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.8618	.9329	.4168	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.7440	.9216	.1819	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.7237	.9346	.1634	10	1

(a) $\alpha = 0.20$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.8974	.9287	.5147	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.8009	.9206	.2707	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.7623	.9303	.2053	10	1

(a) $\alpha = 0.40$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.9323	.9240	.6063	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.8518	.9152	.3802	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.8141	.9197	.2922	10	1

(a) $\alpha = 0.60$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.9532	.9176	.6473	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.8849	.9040	.4493	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.8384	.9012	.3478	10	1

(a) $\alpha = 0.80$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.9701	.9090	.6544	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.9047	.8864	.4374	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.8630	.8735	.3553	10	1

(a) $\alpha = 0.90$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.9757	.8973	.6212	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.9114	.8735	.3890	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.8696	.8623	.2919	10	1

(a) $\alpha = 0.95$

Gross train	Net train	Gross validation	Net validation	Gross test	Net test	AUC-ROC	AUC-PR	R^2	Age of success	Age features
56,697	54,889	28,348	27,391	198,441	192,202	.9753	.8795	.5826	2	1
56,697	22,440	28,348	11,236	198,441	78,462	.9241	.8520	.3377	6	1
56,697	10,031	28,348	5,062	198,441	35,148	.8744	.8396	.2448	10	1

Table B.4.14: Reduction in R-Squared for survival

Age to survive	age 2	age 3	age 6	age 10
R-Squared of the full model	13.99	3.91	0.55	0.14
percentage points reduction in out of sample R^2 when removing				
Financial statement variables and year	7.30	1.94	0.28	0.06
Financial statement variables	7.12	1.96	0.22	0.08
Year	-0.14	-0.01	0.01	-0.06
Worker and dga variables	0.93	0.38	0.04	0.08
Worker variables	0.37	-0.09	-0.18	-0.01
Dga variables	0.07	0.14	-0.02	-0.04
Age size sector area	0.02	0.02	-0.05	-0.03
Age	-0.29	-0.13	0.03	-0.03
Size class (OECD)	-0.11	-0.02	-0.05	0.00
Sector	-0.13	0.08	-0.14	0.06
Region	-0.25	0.09	0.00	0.00
Financial constraints variables	-0.06	0.01	-0.07	-0.05

Table B.4.15: Reduction in R-Squared for economic success (sector-specific cutoffs, $\alpha = 0.2$)

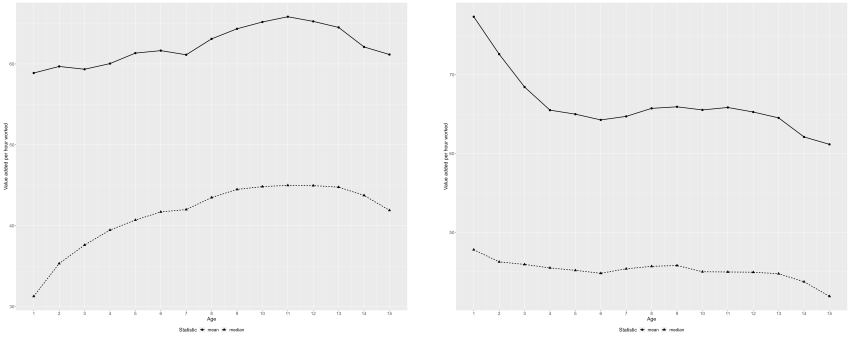
Age to survive	age 2	age 3	age 6	age 10
R-Squared of the full model	61.59	37.57	26.79	20.78
percentage points reduction in out of sample R^2 when removing				
Financial statement variables and year	38.19	19.07	12.04	7.64
Financial statement variables	38.29	19.14	12.00	7.96
Year	0.07	0.02	0.03	-0.10
Worker and dgas variables	0.25	0.35	0.85	1.91
Worker variables	0.07	-0.10	0.15	0.24
Dga variables	-0.00	0.11	-0.18	0.18
Age size sector area	13.27	5.80	4.08	2.65
Age	-0.01	-0.15	-0.13	0.12
Size class (OECD)	0.04	-0.00	-0.01	0.74
Sector	13.46	5.66	4.13	2.48
Region	0.13	-0.07	-0.20	-0.26
Financial constraints variables	-0.08	-0.18	-0.42	-0.09

Table B.4.16: Most important predictors with LASSO for linearized success probabilities

Feature	Coefficient
Constant	74.8318
Year 2021	0.3588
Year 2016	-0.3171
Year 2017	-0.2806
Year 2015	-0.2746
Year 2014	-0.1975
Year 2020	0.1930
Year 2012	0.1780
Year 2018	-0.1672
Year 2011	0.1345
Year 2010	0.1062
Year 2009	0.0679
Year 2013	-0.0570
Region: Zuidoost-Drenthe	-0.0447
Age	-0.0381
R^2	0.8657

B.5 Additional figures

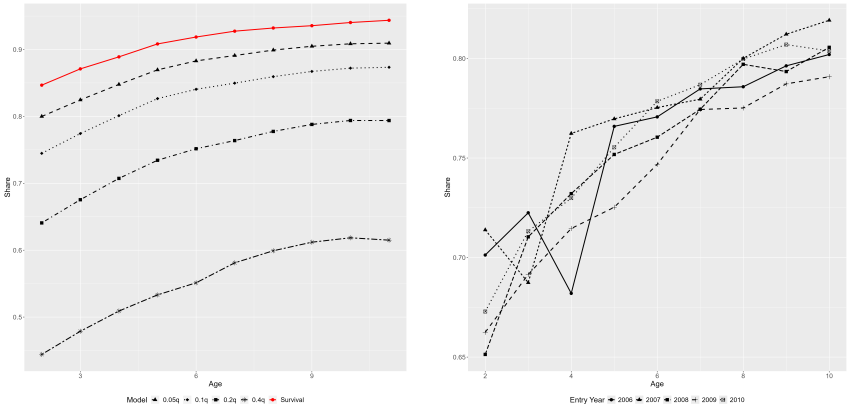
Figure B.5.1: Labor productivity dynamics



(a) All startups

(b) Startups that survived at least 10 years

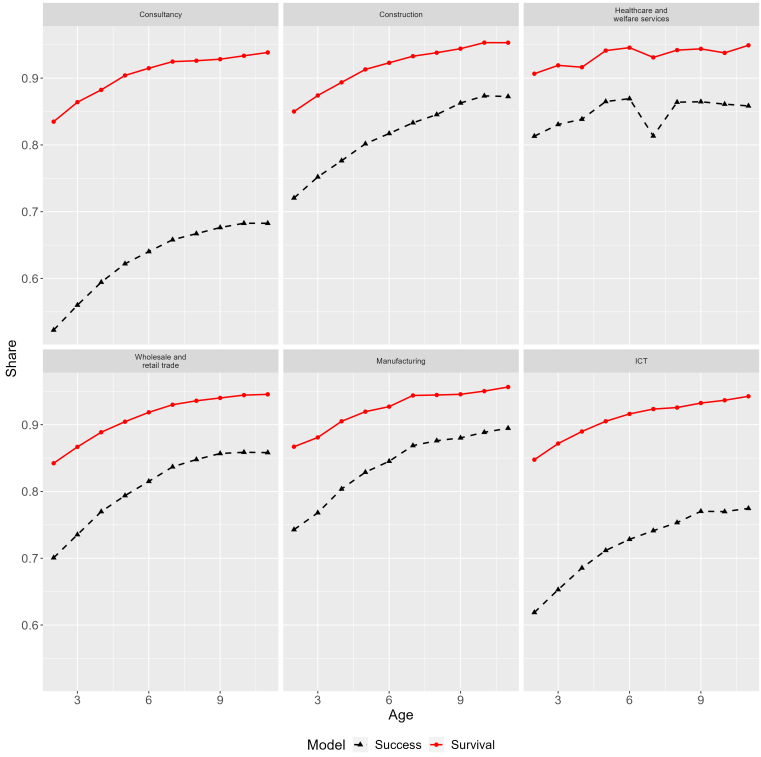
Figure B.5.2: Success at different quantiles and cohorts of entry



(a) Success rates for different cutoffs

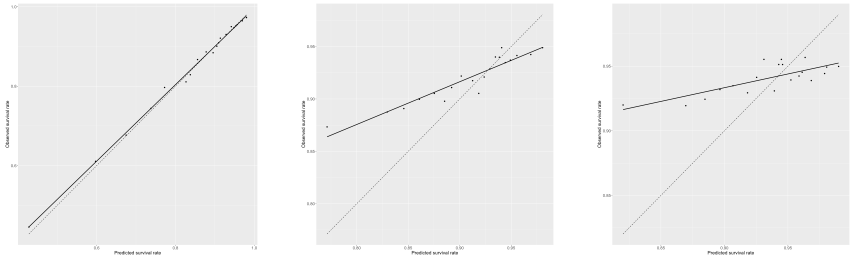
(b) Success rates by cohort of entry

Figure B.5.3: Survival and economic success by industry



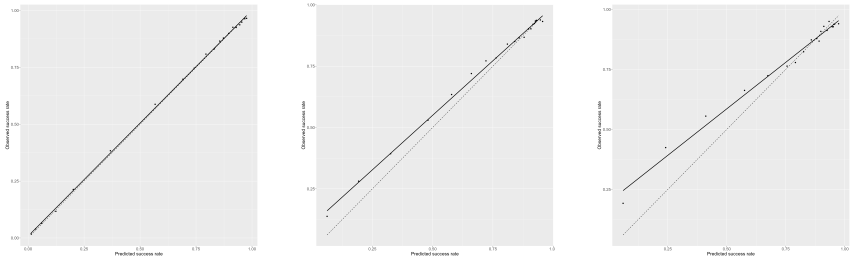
Notes: Success of firm at age a corresponds to survival up to that point, conditional on reaching 150,000€ turnover size (in real terms).

Figure B.5.4: Calibration plots for early-life survival probability



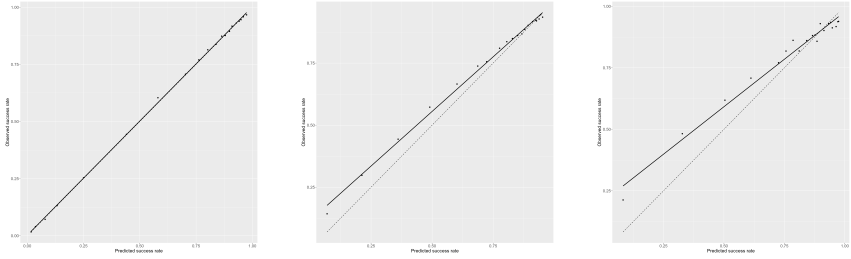
(a) Age 2 **(b)** Age 6 **(c)** Age 10

Figure B.5.5: Calibration plots for early-life success probability (pooled cutoff, $\alpha = 0.2$)



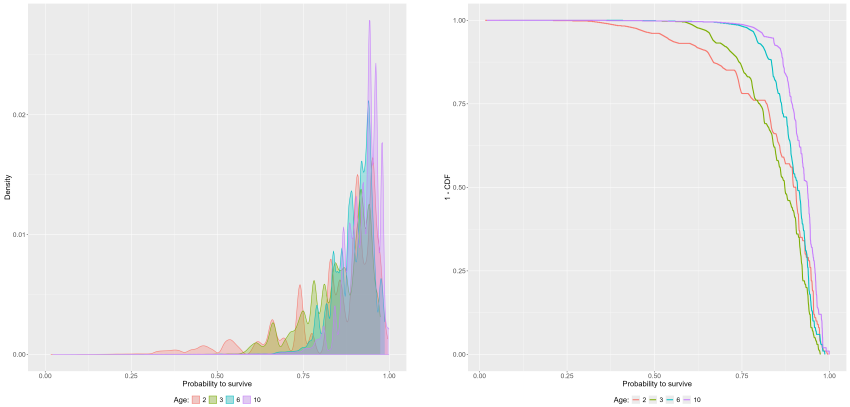
(a) Age 2 **(b)** Age 6 **(c)** Age 10

Figure B.5.6: Calibration plots for early-life success probability (sector-specific cutoffs, $\alpha = 0.2$)



(a) Age 2 **(b)** Age 6 **(c)** Age 10

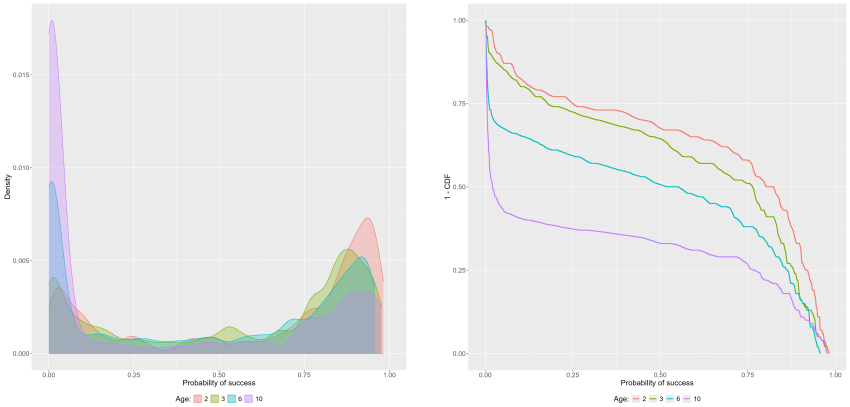
Figure B.5.7: Survival probabilities at various ages



(a) Density **(b)** CCDF

Notes: Densities and CCDF are based on out-of-sample survival probabilities for all testing startups

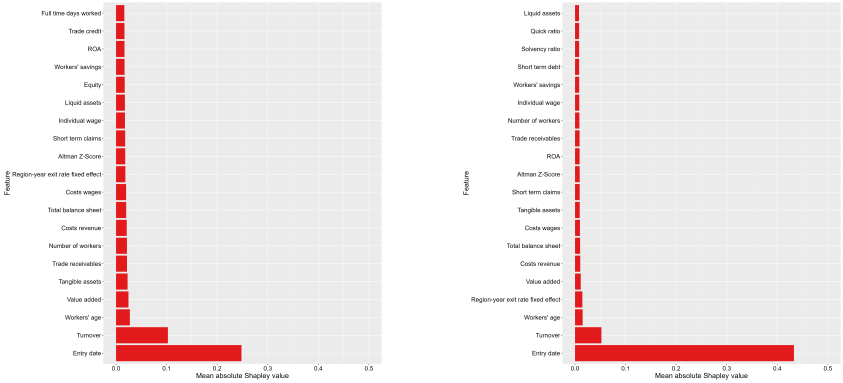
Figure B.5.8: Success type probabilities at various ages (sector-specific cut-offs)



(a) Density **(b)** CCDF

Notes: Densities and CCDFare based on out-of-sample success probabilities for all testing startups

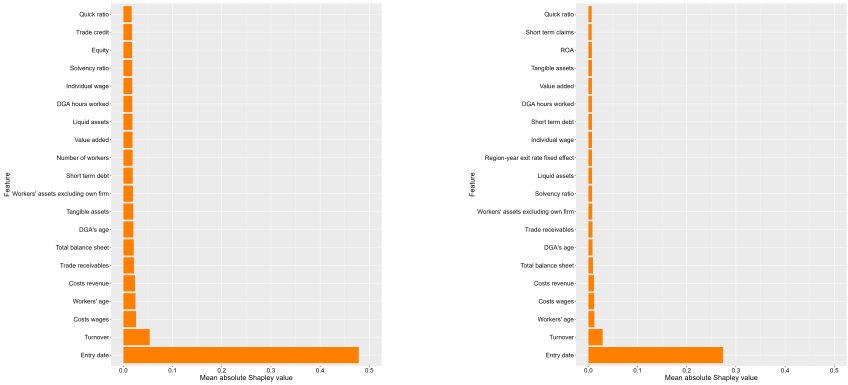
Figure B.5.9: SHAP comparison for success at age 6: top vs. bottom quartile firms



(a) Top 25%

(b) Bottom 25%

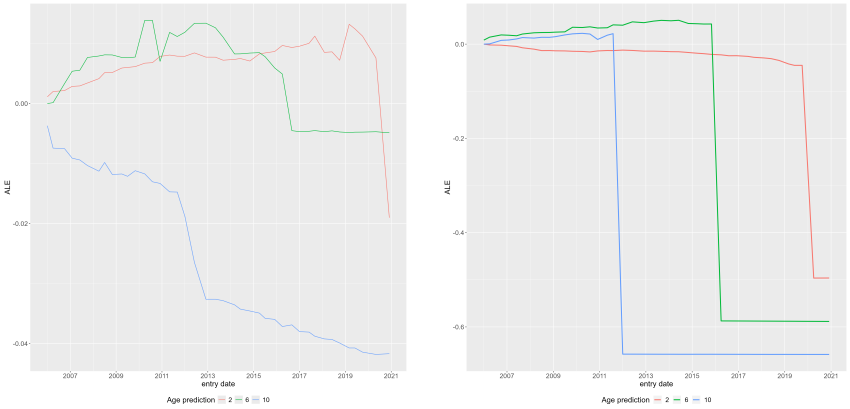
Figure B.5.10: SHAP comparison for success at age 10: top vs. bottom quartile firms



(a) Top 25%

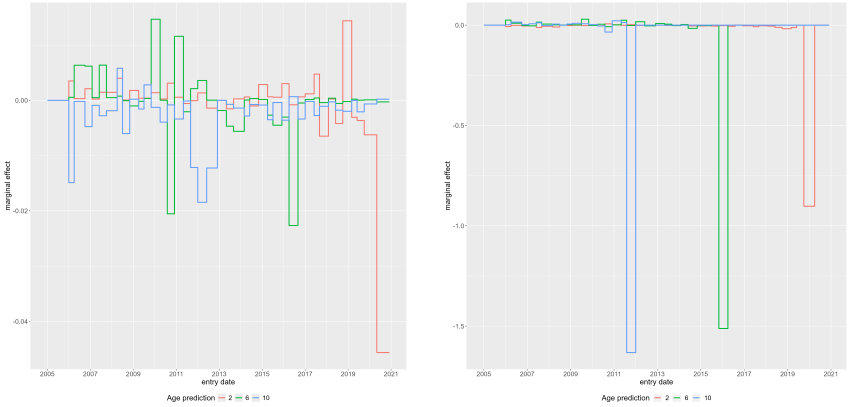
(b) Bottom 25%

Figure B.5.11: ALE plots for entry date at various ages



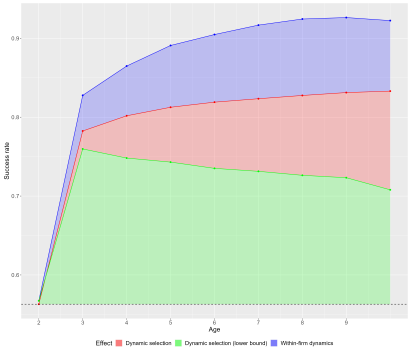
(a) Survival **(b)** Economic success (20p pooled cutoff)

Figure B.5.12: Marginal effects for entry date at various ages

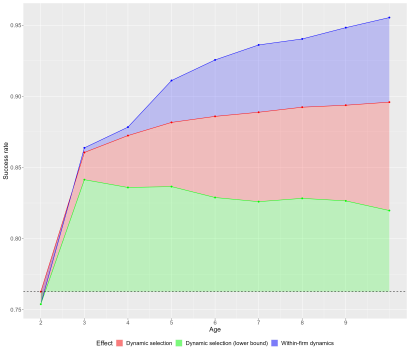


(a) Survival **(b)** Economic success (20p pooled cutoff)

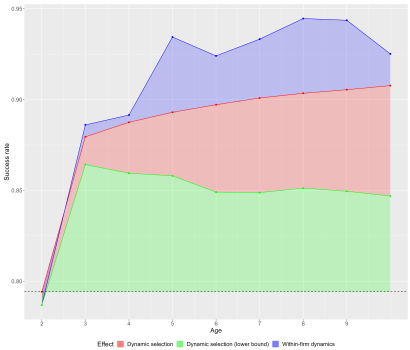
Figure B.5.13: Dynamic selection for economic success by industry



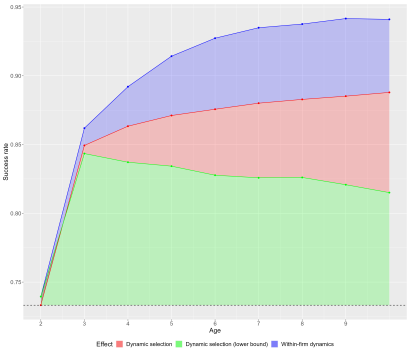
(a) Consultancy



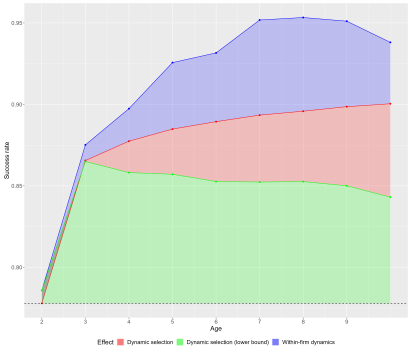
(b) Construction



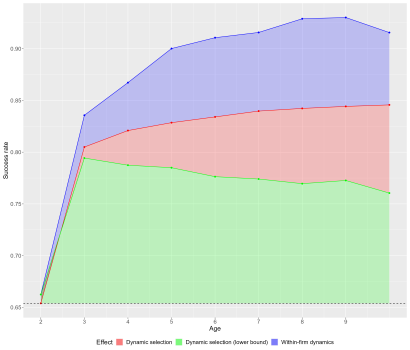
(c) Healthcare and welfare services



(d) Wholesale and retail trade



(e) Manufacturing



(f) ICT

Appendix C

Appendix to Chapter 3

C.1 Additional tables

Table C.1.1: First stage out-of-sample prediction (missing g_1)

Method	AUC-ROC	AUC-PR	F1-Score	R^2	BACC	Num. obs
(a) Employment						
Probit	0.6756	0.1140	0.1205	0.0289	0.6244	176,694
Random forest	0.8271	0.1961	0.1895	0.0894	0.7548	176,694
XGBoost	0.9909	0.9919	0.9725	0.9160	0.9733	176,694
(b) Revenues						
Probit	0.8250	0.5152	0.2852	0.3465	0.7442	197,045
Random forest	0.8905	0.5936	0.3352	0.4026	0.8101	197,045
XGBoost	0.9902	0.9898	0.9714	0.9197	0.9728	197,045

Table C.1.2: First stage out-of-sample prediction (missing g_2)

Method	AUC-ROC	AUC-PR	F1-Score	R^2	BACC	Num. obs
(a) Employment						
Probit	0.6853	0.1760	0.1883	0.0453	0.6310	128,149
Random forest	0.7922	0.2481	0.2520	0.1024	0.7181	128,149
XGBoost	0.9852	0.9893	0.9637	0.8803	0.9649	128,149
(b) Revenues						
Probit	0.8150	0.4909	0.2929	0.2911	0.7268	144,098
Random forest	0.8743	0.5610	0.3570	0.3430	0.7900	144,098
XGBoost	0.9885	0.9897	0.9666	0.9003	0.9679	144,098

Table C.1.3: First-stage regression: probit model

	Missing g_n					
	Revenues			Employees		
	(g_1)	(g_2)	(g_3)	(g_1)	(g_2)	(g_3)
Age	-0.040^{***}	-0.051^{***}	-0.073^{***}	-0.002	-0.012^{**}	-0.006
	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Size (log)	-0.602^{***}	-0.596^{***}	-0.609^{***}	-0.319^{***}	-0.325^{***}	-0.312^{***}
	(0.002)	(0.003)	(0.003)	(0.006)	(0.006)	(0.006)
Consolidated accounts	0.033^{***}	0.042^{***}	0.048^{***}	0.089^{***}	0.091^{***}	0.088^{***}
	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.004)
Patents	-0.048^{***}	-0.055^{***}	-0.061^{***}	-0.019^{***}	-0.027^{***}	-0.035^{***}
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Trademarks	-0.066^{***}	-0.071^{***}	-0.076^{***}	-0.025^{***}	-0.031^{***}	-0.038^{***}
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Share past missing	-0.130^{***}	-0.178^{***}	-0.259^{***}	-0.302^{***}	-0.340^{***}	-0.314^{***}
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Training obs.	784,618	573,257	371,028	703,119	509,226	325,165

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: coefficients are standardized as $\beta_x^* = \beta_x \sigma_x$ to improve comparability. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent estimation.

Table C.1.4: Second-stage regressions: one-year models comparison

One-year growth rate (g_1)						
	Revenues			Employment		
	(a)	(b)	(c)	(a)	(b)	(c)
Intercept	-15.448^{***}	-1028.702^{***}	-21.052^{***}	-2.629^{***}	-274.595^{***}	2.403^{***}
	(0.595)	(153.860)	(1.375)	(0.457)	(14.958)	(0.632)
Age	-0.000^{***}	-0.008^{***}	-0.000^{***}	-0.000^{***}	-0.001^{***}	-0.000^{***}
	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
Size (log)	-0.009^{***}	-0.677^{***}	-0.010^{***}	-0.015^{***}	-0.419^{***}	-0.013^{***}
	(0.002)	(0.103)	(0.002)	(0.001)	(0.022)	(0.001)
Consolidated accounts	0.039^{***}	4.592^{***}	0.048^{***}	0.044^{***}	2.127^{***}	0.035^{***}
	(0.008)	(0.695)	(0.010)	(0.006)	(0.118)	(0.006)
Patents	0.013^{***}	-0.012^{***}	0.013^{***}	0.011^{***}	-0.131^{***}	0.012^{***}
	(0.001)	(0.003)	(0.001)	(0.001)	(0.008)	(0.001)
Trademarks	0.013^{***}	-0.019^{***}	0.012^{***}	0.012^{***}	-0.178^{***}	0.013^{***}
	(0.002)	(0.004)	(0.002)	(0.001)	(0.010)	(0.001)
IMR		3.197^{***}	0.009^{***}		2.002^{***}	-0.011^{***}
		(0.486)	(0.002)		(0.110)	(0.001)
PCA Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
First Stage		Probit	XGB		Probit	XGB
Second Stage	OLS	OLS	OLS	OLS	OLS	OLS
R^2	0.0148	0.0700	0.0149	0.0185	0.0260	0.0192
RMSE	0.1886	0.1833	0.1886	0.1385	0.1379	0.1384
Obs. 1st Stage		784,618	784,618		703,119	703,119
Obs. 2nd Stage	202,755	202,755	202,755	197,758	197,758	197,758

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: PCA controls include lagged predictors condensed into five principal components. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent (HC3) estimation.

Table C.1.5: Second-stage regressions: two-year models comparison

Two-year growth rate (g_2)						
	Revenues			Employment		
	(a)	(b)	(c)	(a)	(b)	(c)
Intercept	-36.670^{***}	-1236.931^{***}	-44.383^{***}	-15.366^{***}	-556.759^{***}	-7.686^{***}
	(1.000)	(263.666)	(1.967)	(0.834)	(24.542)	(1.069)
Age	-0.001^{***}	-0.008^{***}	-0.001^{***}	-0.001^{***}	-0.002^{***}	-0.001^{***}
	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
Size (log)	-0.006^{**}	-0.443^{***}	-0.006^{**}	-0.024^{***}	-0.436^{***}	-0.021^{***}
	(0.002)	(0.098)	(0.002)	(0.001)	(0.019)	(0.001)
Consolidated accounts	0.052^{***}	3.048^{***}	0.059^{***}	0.061^{***}	2.146^{***}	0.049^{***}
	(0.011)	(0.663)	(0.011)	(0.007)	(0.098)	(0.007)
Patents	0.020^{***}	-0.048^{***}	0.020^{***}	0.019^{***}	-0.174^{***}	0.021^{***}
	(0.002)	(0.014)	(0.002)	(0.001)	(0.009)	(0.001)
Trademarks	0.017^{***}	-0.005	0.017^{***}	0.022^{***}	-0.204^{***}	0.024^{***}
	(0.002)	(0.004)	(0.002)	(0.001)	(0.010)	(0.001)
IMR		2.172^{***}	0.008^{***}		2.062^{***}	-0.015^{***}
		(0.477)	(0.002)		(0.093)	(0.001)
PCA Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
First Stage		Probit	XGB		Probit	XGB
Second Stage	OLS	OLS	OLS	OLS	OLS	OLS
R^2	0.0287	0.0528	0.0288	0.0348	0.0452	0.0356
RMSE	0.2253	0.2224	0.2252	0.1759	0.1749	0.1758
Obs. 1st Stage		573,257	573,257		509,226	509,226
Obs. 2nd Stage	140,509	140,509	140,509	135,668	135,668	135,668

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: PCA controls include lagged predictors condensed into five principal components. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent (HC3) estimation.

Table C.1.6: Second-stage regressions: models based on Random Forest first-stage

	n-year growth rate (g_n)					
	Revenues			Employees		
	(g_1)	(g_2)	(g_3)	(g_1)	(g_2)	(g_3)
Intercept	-18.482^{***}	-46.361^{***}	-76.461^{***}	7.280^{***}	3.725^{**}	1.284
	(1.129)	(2.015)	(4.654)	(0.674)	(1.266)	(2.454)
Age	-0.000^{***}	-0.001^{***}	-0.001^{***}	-0.000^{***}	-0.001^{***}	-0.001^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Size (log)	-0.009^{***}	-0.007^{***}	-0.006^{**}	-0.010^{***}	-0.017^{***}	-0.022^{***}
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Consolidated accounts	0.043^{***}	0.062^{***}	0.086^{***}	0.022^{***}	0.031^{***}	0.065^{***}
	(0.009)	(0.011)	(0.015)	(0.006)	(0.007)	(0.010)
Patents	0.013^{***}	0.019^{***}	0.024^{***}	0.013^{***}	0.024^{***}	0.031^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Trademarks	0.012^{***}	0.016^{***}	0.018^{***}	0.014^{***}	0.027^{***}	0.036^{***}
	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
IMR	0.007^{***}	0.015^{***}	0.020^{***}	-0.032^{***}	-0.050^{***}	-0.060^{***}
	(0.002)	(0.003)	(0.005)	(0.002)	(0.003)	(0.004)
PCA Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
First Stage	RF	RF	RF	RF	RF	RF
Second Stage	OLS	OLS	OLS	OLS	OLS	OLS
R^2	0.0148	0.0289	0.0414	0.0208	0.0381	0.0504
RMSE	0.1886	0.2252	0.2461	0.1383	0.1756	0.1963
Obs. 1st Stage	784,618	573,257	371,028	703,119	509,226	325,165
Obs. 2nd Stage	202,755	140,509	82,861	197,758	135,668	79,026

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Note: PCA controls include lagged predictors condensed into five principal components. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent (HC3) estimation.

Table C.1.7: Second-stage regressions: models based on XGBoost-generalized first-stage

	n-year growth rate (g_n)					
	Revenues			Employees		
	(g_1)	(g_2)	(g_3)	(g_1)	(g_2)	(g_3)
Intercept	-15.756^{***}	-36.652^{***}	-62.988^{***}	-2.759^{***}	-13.512^{***}	-16.471^{***}
	(0.582)	(1.071)	(3.662)	(0.461)	(0.836)	(2.028)
Age	-0.000^{***}	-0.001^{***}	-0.001^{***}	-0.000^{***}	-0.001^{***}	-0.001^{***}
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Size (log)	-0.009^{***}	-0.006^{**}	-0.005^{*}	-0.014^{***}	-0.022^{***}	-0.028^{***}
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Consolidated accounts	0.044^{***}	0.052^{***}	0.076^{***}	0.041^{***}	0.051^{***}	0.089^{***}
	(0.009)	(0.011)	(0.015)	(0.006)	(0.007)	(0.010)
Patents	0.013^{***}	0.020^{***}	0.026^{***}	0.011^{***}	0.019^{***}	0.025^{***}
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Trademarks	0.013^{***}	0.017^{***}	0.019^{***}	0.013^{***}	0.023^{***}	0.028^{***}
	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)
IMR	0.005^{**}	0.000	0.003	-0.005^{***}	-0.012^{***}	-0.009^{***}
	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
PCA Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
First Stage	XGB	XGB	XGB	XGB	XGB	XGB
Second Stage	OLS	OLS	OLS	OLS	OLS	OLS
R^2	0.0149	0.0287	0.0414	0.0187	0.0362	0.0477
RMSE	0.1886	0.2253	0.2461	0.1384	0.1757	0.1963
Obs. 1st Stage	784,618	573,257	371,028	703,119	509,226	325,165
Obs. 2nd Stage	202,755	140,509	82,861	197,758	135,668	79,026

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Note: PCA controls include lagged predictors condensed into five principal components. All regressions include industry, regional, and year fixed effects. Standard errors are based on heteroscedasticity-consistent (HC3) estimation.

Table C.1.8: Second-stage regressions: coefficients of size with two-way fixed effects model

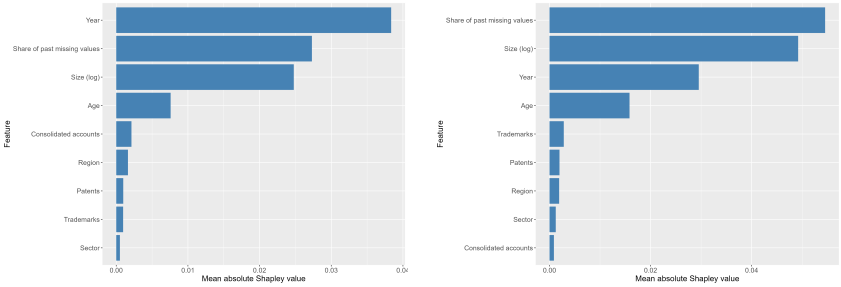
n-year growth rate (g_n)						
	Revenues			Employees		
First stage	(g_1)	(g_2)	(g_3)	(g_1)	(g_2)	(g_3)
No first stage	-0.229^{**}	-0.285^{**}	-0.327^{**}	-0.354^{**}	-0.482^{***}	-0.511^{***}
	(0.032)	(0.024)	(0.026)	(0.043)	(0.035)	(0.009)
Probit	-0.229^{**}	-0.284^{**}	-0.326^{**}	-0.354^{**}	-0.482^{***}	-0.512^{***}
	(0.032)	(0.024)	(0.026)	(0.043)	(0.035)	(0.009)
Random Forest	-0.231^{**}	-0.287^{**}	-0.328^{**}	-0.352^{**}	-0.478^{***}	-0.507^{***}
	(0.032)	(0.024)	(0.027)	(0.044)	(0.036)	(0.009)
XGBoost	-0.229^{**}	-0.284^{**}	-0.327^{**}	-0.354^{**}	-0.481^{***}	-0.510^{***}
	(0.032)	(0.024)	(0.026)	(0.044)	(0.036)	(0.009)
Obs. 1st Stage	784, 618	573, 257	371, 028	703, 119	509, 226	325, 165
Obs. 2nd Stage	202, 755	140, 509	82, 861	197, 758	135, 668	79, 026

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

The models estimate two-way fixed effects panel regressions, incorporating both firm-level and time fixed effects. Variables like trademarks, patents and consolidated accounts are excluded due to collinearity with fixed effects. Standard errors are clustered at both firm and time levels for robust inference.

C.2 Additional figures

Figure C.2.1: Shapley values for Probit first stage prediction (missing g_3)



(a) Revenues **(b)** Employment

Bibliography

Abbasi, Raza Abid et al. (2019). "Short term load forecasting using XGBoost". In: *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33*. Springer, pp. 1120–1131.

Acemoglu, Daron et al. (2018). "Innovation, reallocation, and growth". In: *American Economic Review* 108.11, pp. 3450–3491.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer (2015). "Identification properties of recent production function estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Adalet McGowan, Müge, Dan Andrews, and Valentine Millot (2018). "The walking dead? Zombie firms and productivity performance in OECD countries". In: *Economic Policy* 33.96, pp. 685–736.

Agarwal, Rajshree and David B Audretsch (2001). "Does entry size matter? The impact of the life cycle and technology on firm survival". In: *The Journal of Industrial Economics* 49.1, pp. 21–43.

Aghion, Philippe, Ufuk Akcigit, and Peter Howitt (2014). "What do we learn from Schumpeterian growth theory?" In: *Handbook of economic growth*. Vol. 2. Elsevier, pp. 515–563.

Aghion, Philippe, Antonin Bergeaud, et al. (2019). "Coase Lecture - The Inverted-U Relationship Between Credit Access and Productivity Growth". In: *Economica* 86.341, pp. 1–31.

Ahrens, Achim, Christian B Hansen, and Mark E Schaffer (2019). "lassopack: Model selection and prediction with regularized regression in Stata". In: *arXiv preprint arXiv:1901.05397*.

— (2020). "lassopack: Model selection and prediction with regularized regression in Stata". In: *The Stata Journal* 20.1, pp. 176–235.

Akcigit, Ufuk and Sina T Ates (2021). "Ten facts on declining business dynamism and lessons from endogenous growth theory". In: *American Economic Journal: Macroeconomics* 13.1, pp. 257–298.

Alaka, Hafiz A et al. (2018). "Systematic review of bankruptcy prediction models: Towards a framework for tool selection". In: *Expert Systems with Applications* 94, pp. 164–184.

Altman, Edward I (1968). "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy". In: *The Journal of Finance* 23.4, pp. 589–609.

Altman, Edward I et al. (2000). "Predicting financial distress of companies: revisiting the Z-score and ZETA models". In: *Stern School of Business, New York University*, pp. 9–12.

Andrews, Dan, Muge Adalet McGowan, Valentine Millot, et al. (2017). *Confronting the zombies: Policies for productivity revival*. Tech. rep. OECD Publishing.

Andrews, Dan and Filippos Petroulakis (Feb. 2019). *Breaking the shackles: Zombie firms, weak banks and depressed restructuring in Europe*. Working Paper Series 2240. European Central Bank.

Apley, Daniel W and Jingyu Zhu (2020). "Visualizing the effects of predictor variables in black box supervised learning models". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.4, pp. 1059–1086. DOI: https://doi.org/10.1111/rssb.12377.

Athey, S. (2017). "The Impact of Machine Learning on Economics". In: *In Economics of Artificial Intelligence, NBER Chapters*.

Athey, Susan (2018). "The impact of machine learning on economics". In: *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Audretsch, David B and Talat Mahmood (1994). "The rate of hazard confronting new firms and plants in US manufacturing". In: *Review of Industrial organization* 9, pp. 41–56.

— (1995). "New firm survival: new results using a hazard function". In: *The review of economics and statistics*, pp. 97–103.

Bajgar, Matej et al. (2020). "Coverage and representativeness of Orbis data". In.

Banerjee, Ryan and Boris Hofmann (2018). "The rise of zombie firms: causes and consequences". In: *BIS Quarterly Review Spetember*.

Bank of England (2013). *Inflation Report*. Tech. rep. Bank of England, London.

Bank of Korea (2013). *Financial Stability Report*. Tech. rep. Bank of Korea, Seoul.

Bargagli-Stoffi, Falco J, Fabio Incerti, et al. (2024a). "Machine learning for zombie hunting: predicting distress from firms' accounts and missing values". In: *Industrial and Corporate Change* 33.5, pp. 1063–1097.

— (2024b). "Machine learning for zombie hunting: predicting distress from firms' accounts and missing values". In: *Industrial and Corporate Change* 33.5, pp. 1063–1097.

Bargagli-Stoffi, Falco J, Jan Niederreiter, and Massimo Riccaboni (2021). "Supervised learning for the prediction of firm dynamics". In: *Data science for economics and finance*. Springer, Cham, pp. 19–41.

Bargagli-Stoffi, Falco J., Gustavo Cevolani, and Giorgio Gnecco (2022). "Simple models in complex worlds: Occam's razor and statistical learning theory". In: *Minds and Machines*. in press.

Bargagli-Stoffi, Falco Joannes, Gustavo Cevolani, and Giorgio Gnecco (2020). "Should Simplicity Be Always Preferred to Complexity in Supervised Machine Learning?" In: *6th International Conference on machine Learning, Optimization Data science (LOD2020)*. Lecture Notes in Computer Science. Springer.

Bartelsman, Eric J and Mark Doms (2000). "Understanding productivity: Lessons from longitudinal microdata". In: *Journal of Economic literature* 38.3, pp. 569–594.

Behr, Andreas and Jurij Weinblat (2017). "Default patterns in seven EU countries: A random forest approach". In: *International Journal of the Economics of Business* 24.2, pp. 181–222.

Bellman, Richard E. (1961). *Adaptive Control Processes*. Princeton University Press.

Belloni, Alexandre, Victor Chernozhukov, Christian Hansen, et al. (2016). "Inference in high-dimensional panel models with an application to gun control". In: *Journal of Business & Economic Statistics* 34.4, pp. 590–605.

Belloni, Alexandre, Victor Chernozhukov, and Ying Wei (2016). "Post-selection inference for generalized linear models with many controls". In: *Journal of Business & Economic Statistics* 34.4, pp. 606–619.

Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz (2021). "A comparative analysis of gradient boosting algorithms". In: *Artificial Intelligence Review* 54, pp. 1937–1967.

Black, Fischer and Myron Scholes (1973). "The pricing of options and corporate liabilities". In: *Journal of political economy* 81.3, pp. 637–654.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen (2016). *Management as a Technology?* Vol. 22327. National Bureau of Economic Research Cambridge, MA.

Bonfim, Diana et al. (2020). "On-site inspecting zombie lending". In: *Available at SSRN 3530574*.

Brédart, Xavier (2014). "Bankruptcy prediction model using neural networks". In: *Accounting and Finance Research* 3.2, pp. 124–128.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24, pp. 123–140.

— (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Breiman, Leo et al. (1984). "Classification and regression trees". In: *Belmont, CA: Wadsworth & Brooks*.

Brodersen, Kay Henning et al. (2010). "The balanced accuracy and its posterior distribution". In: *2010 20th International Conference on Pattern Recognition*. IEEE, pp. 3121–3124.

Buckmann, Marcus, Andreas Joseph, and Helena Robertson (2021). "An interpretable machine learning workflow with an application to economic forecasting". In.

Bugamelli, Matteo et al. (Jan. 2018). *Productivity growth in Italy: a tale of a slow-motion change*. Questioni di Economia e Finanza (Occasional Papers) 422. Bank of Italy, Economic Research and International Relations Area.

Buldyrev, Sergey et al. (2020). *The Rise and Fall of Business Firms: A Stochastic Framework on Innovation, Creative Destruction and Growth*. Cambridge University Press.

Caballero, Ricardo J, Takeo Hoshi, and Anil K Kashyap (2008). "Zombie lending and depressed restructuring in Japan". In: *American Economic Review* 98.5, pp. 1943–77.

Calligaris, Sara, Massimo Del Gatto, Fadi Hassan, Gianmarco I P Ottaviano, et al. (Sept. 2018). "The productivity puzzle and misallocation: an Italian perspective". In: *Economic Policy* 33.96, pp. 635–684.

Calligaris, Sara, Massimo Del Gatto, Fadi Hassan, Gianmarco IP Ottaviano, et al. (2016). *Italy's productivity conundrum. a study on resource misallocation in Italy*. Tech. rep. Directorate General Economic and Financial Affairs (DG ECFIN), European Commission.

Cameron, AC (2005). "Microeconometrics: methods and applications". In: *Cambridge University*.

Capasso, Marco, Tania Treibich, and Bart Verspagen (2015). "The medium-term effect of R&D on firm growth". In: *Small Business Economics* 45, pp. 39–62.

Cascarano, Michele, Filippo Natoli, and Andrea Petrella (2025). "Entry, exit, and market structure in a changing climate". In: *European Economic Review*, p. 105027.

Cavallari, Lilia, Simone Romano, and Paolo Naticchioni (2021). "The original sin: Firms' dynamics and the life-cycle consequences of economic conditions at birth". In: *European Economic Review* 138, p. 103844.

Chen, Jiahua and Zehua Chen (2008). "Extended Bayesian information criteria for model selection with large model spaces". In: *Biometrika* 95.3, pp. 759–771.

Chen, Tianqi and Carlos Guestrin (2016). "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Chernozhukov, Victor et al. (2018). *Double/debiased machine learning for treatment and structural parameters*.

Chipman, Hugh A, Edward I George, Robert E McCulloch, et al. (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Coad, Alex (2009). *The growth of firms: A survey of theories and empirical evidence*. Edward Elgar Publishing.

Coad, Alex et al. (2013). "Growth paths and survival chances: An application of Gambler's Ruin theory". In: *Journal of Business Venturing* 28.5, pp. 615–632.

Cooper, Arnold C, F Javier Gimeno-Gascon, and Carolyn Y Woo (1994). "Initial human and financial capital as predictors of new venture performance". In: *Journal of business venturing* 9.5, pp. 371–395.

Cravino, Javier and Andrei A Levchenko (2017). "Multinational firms and international business cycle transmission". In: *The Quarterly Journal of Economics* 132.2, pp. 921–962.

— (Nov. 2016). "Multinational Firms and International Business Cycle Transmission". In: *The Quarterly Journal of Economics* 132.2, pp. 921–962.

Cucculelli, Marco and Valentina Peruzzi (2020). "Post-crisis firm survival, business model changes, and learning: evidence from the Italian manufacturing industry". In: *Small Business Economics* 54.2, pp. 459–474. DOI: https://doi.org/10.1007/s11187-018-0044-2.

Davies, Lily, Mark Kattenberg, and Benedikt Vogt (2023). *Predicting Firm Exits with Machine Learning: Implications for Selection into COVID-19 Support and Productivity Growth*. Tech. rep. CPB Netherlands Bureau for Economic Policy Analysis.

De Martiis, Angela, Thomas LA Heil, and Franziska J Peter (2020). "Are you a Zombie? A Supervised Learning Method to Classify Unviable Firms and Identify the Determinants". In.

Decker, Ryan A et al. (2016). "Declining business dynamism: What we know and the way forward". In: *American Economic Review* 106.5, pp. 203–207.

Dinlersoz, Emin et al. (2018). *Leverage over the life cycle and implications for firm growth and shock responsiveness*. Tech. rep. National Bureau of Economic Research.

Dunne, Timothy, Mark J Roberts, and Larry Samuelson (1989). "The growth and failure of US manufacturing plants". In: *The Quarterly Journal of Economics* 104.4, pp. 671–698.

Evans, David S (1987a). "Tests of alternative theories of firm growth". In: *journal of political economy* 95.4, pp. 657–674.

— (1987b). "The relationship between firm growth, size, and age: Estimates for 100 manufacturing industries". In: *The journal of industrial economics*, pp. 567–581.

Fairlie, Robert et al. (2025). "The Effects of Local Taxes and Incentives on Entrepreneurship: Evidence from the Universe of US Startups". In: *Available at SSRN 5262559*.

Farinas, Jose C and Lourdes Moreno (2000). "Firms' growth, size and age: A nonparametric approach". In: *Review of Industrial organization* 17, pp. 249–265.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8, pp. 861–874.

Fazzari, Steven M., R. Glenn Hubbard, and Bruce C. Petersen (1988). "Financing Constraints and Corporate Investment". In: *Brookings Papers on Economic Activity* 19.1, pp. 141–206.

Ferrando, Annalisa and Alessandro Ruggieri (2018). "Financial constraints and productivity: Evidence from euro area companies". In: *International Journal of Finance & Economics* 23.3, pp. 257–282.

Freund, Yoav, Robert E Schapire, et al. (1996). "Experiments with a new boosting algorithm". In: *icml*. Vol. 96. Citeseer, pp. 148–156.

Fuertes-Callén, Yolanda, Beatriz Cuellar-Fernández, and Carlos Serrano-Cinca (2022). "Predicting startup survival using first years financial statements". In: *Journal of Small Business Management* 60.6, pp. 1314–1350. DOI: https://doi.org/10.1080/00472778.2020.1750302.

Gal, Peter N (2013). "Measuring total factor productivity at the firm level using OECD-ORBIS". In.

Garnsey, E (1998). "A theory of the early growth of the firm". In: *Industrial and corporate change* 7.3, pp. 523–556. DOI: https://doi.org/10.1111/rssb.12377.

Geroski, Paul A, José Mata, and Pedro Portugal (2010). "Founding conditions and the survival of new firms". In: *Strategic Management Journal* 31.5, pp. 510–529.

Gibrat, Robert (1931). "Les inégalités économiques: applications aux inégalités des richesses, à la concentration des entreprises... d'une loi nouvelle, la loi de l'effet proportionnel". In: *Paris: Librairie du Recueil Sirey*.

Gimeno, Javier et al. (1997). "Survival of the fittest? Entrepreneurial human capital and the persistence of underperforming firms". In: *Administrative science quarterly*, pp. 750–783.

Glaeser, Stephen and James D Omartian (2022). "Public firm presence, financial reporting, and the decline of US manufacturing". In: *Journal of Accounting Research* 60.3, pp. 1085–1130.

Gopinath, Gita et al. (2017). "Capital allocation and productivity in South Europe". In: *The Quarterly Journal of Economics* 132.4, pp. 1915–1967.

Gourinchas, Pierre-Olivier et al. (2020). *Covid-19 and SME failures*. Vol. 27877. 1. National Bureau of Economic Research Cambridge, MA.

Grazzi, Marco and Daniele Moschella (2018). "Small, young, and exporters: New evidence on the determinants of firm growth". In: *Journal of Evolutionary Economics* 28.1, pp. 125–152.

Greenhalgh, Christine and Mark Rogers (2010). "Innovation, intellectual property, and economic growth". In: *Innovation, intellectual property, and economic growth*. Princeton University Press.

Gumus, Mesut and Mustafa S Kiran (2017). "Crude oil price forecasting using XGBoost". In: *2017 International conference on computer science and engineering (UBMK)*. IEEE, pp. 1100–1103.

Hadlock, Charles J and Joshua R Pierce (2010). "New evidence on measuring financial constraints: Moving beyond the KZ index". In: *The Review of Financial Studies* 23.5, pp. 1909–1940.

Hall, Bronwyn et al. (2014). "The choice between formal and informal intellectual property: a review". In: *Journal of Economic Literature* 52.2, pp. 375–423.

Hall, Bronwyn H (1986). *The relationship between firm size and firm growth in the US manufacturing sector*.

Haltiwanger, John, Ron S Jarmin, and Javier Miranda (2011). "Who creates jobs? Small vs. large vs. young". In: *NBER working paper* 16300.

Hanley, James A and Barbara J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Haveman, Heather A (1995). "The demographic metabolism of organizations: Industry dynamics, turnover, and tenure distributions". In: *Administrative Science Quarterly*, pp. 586–618.

He, Yulei et al. (2010). "Multiple imputation in a large-scale complex survey: a practical guide". In: *Statistical methods in medical research* 19.6, pp. 653–670.

Heckman, James J (1979). "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society*, pp. 153–161.

Helfat, Constance E (2007). *Stylized facts, empirical research and theory development in management*.

Hernández, Belinda et al. (2018). "Bayesian additive regression trees using Bayesian model averaging". In: *Statistics and Computing* 28.4, pp. 869–890.

Hirukawa, Masayuki et al. (2023). "DS-HECK: double-lasso estimation of Heckman selection model". In: *Empirical Economics* 64.6, pp. 3167–3195.

Hosaka, Tadaaki (2019). "Bankruptcy prediction using imaged financial ratios and convolutional neural networks". In: *Expert systems with Applications* 117, pp. 287–299.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3, pp. 651–674.

Hsieh, Chang-Tai and Peter J Klenow (2014). "The life cycle of plants in India and Mexico". In: *The Quarterly Journal of Economics* 129.3, pp. 1035–1084.

Jin, Linlin et al. (2017). "Entrepreneurial team composition characteristics and new venture performance: A meta–analysis". In: *Entrepreneurship theory and practice* 41.5, pp. 743–771.

Josse, Julie et al. (2019). "On the consistency of supervised learning with missing values". In: *arXiv preprint arXiv:1902.06931*.

Jovanovic, Boyan (1982). "Selection and the Evolution of Industry". In: *Econometrica: Journal of the econometric society*, pp. 649–670.

Kalecki, Michał (1945). "On the Gibrat distribution". In: *Econometrica: Journal of the Econometric Society*, pp. 161–170.

Kalemli-Ozcan, Sebnem et al. (2015). *How to construct nationally representative firm level data from the ORBIS global database*. Tech. rep. National Bureau of Economic Research.

Kalemli-Özcan, Şebnem, Luc Laeven, and David Moreno (2022). "Debt overhang, rollover risk, and corporate investment: Evidence from the European crisis". In: *Journal of the European Economic Association* 20.6, pp. 2353–2395.

Kalemli-Özcan, Şebnem, Bent E Sørensen, et al. (2024). "How to Construct Nationally Representative Firm-Level Data from the Orbis Global Database: New Facts on SMEs and Aggregate Implications for Industry Concentration". In: *American Economic Journal: Macroeconomics* 16.2, pp. 353–374.

Kapelner, Adam and Justin Bleich (2015). "Prediction with missing data via Bayesian additive regression trees". In: *Canadian Journal of Statistics* 43.2, pp. 224–239.

— (2016). "bartMachine: Machine Learning with Bayesian Additive Regression Trees". In: *Journal of Statistical Software, Articles* 70.4, pp. 1–40. ISSN: 1548-7660. DOI: 10.18637/jss.v070.i04. URL: https://www.jstatsoft.org/v070/i04.

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo (2016). "Examples are not enough, learn to criticize! criticism for interpretability". In: *Advances in Neural Information Processing Systems*, pp. 2280–2288.

Kleinberg, Jon et al. (2015). "Prediction policy problems". In: *American Economic Review* 105.5, pp. 491–95.

Lee, Kun Chang, Ingoo Han, and Youngsig Kwon (1996). "Hybrid neural network models for bankruptcy predictions". In: *Decision Support Systems* 18.1, pp. 63–72.

Lee, Kwonsang, Falco J Bargagli-Stoffi, and Francesca Dominici (2020). "Causal rule ensemble: Interpretable inference of heterogeneous treatment effects". In: *arXiv preprint arXiv:2009.09036*.

Li, Mingqi, Xiaoyang Fu, and Dongdong Li (2020). "Diabetes prediction based on XGBoost algorithm". In: *IOP conference series: materials science and engineering*. Vol. 768. 7. IOP Publishing.

Linero, Antonio R (2018). "Bayesian regression trees for high-dimensional prediction and variable selection". In: *Journal of the American Statistical Association* 113.522, pp. 626–636.

Linero, Antonio R and Yun Yang (2018). "Bayesian regression tree ensembles that adapt to smoothness and sparsity". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.5, pp. 1087–1110.

Linn, Matthew and Daniel Weagley (2019). "Estimating Financial Constraints with Machine Learning". In: *Available at SSRN 3375048*.

Little, Roderick JA and Donald B Rubin (2019). *Statistical analysis with missing data*. Vol. 793. John Wiley & Sons.

Liu, Grace (2020). "Data quality problems troubling business and financial researchers: A literature review and synthetic analysis". In: *Journal of Business & Finance Librarianship* 25.3-4, pp. 315–371.

Loh, Wei-Yin (2002). "Regression tress with unbiased variable selection and interaction detection". In: *Statistica sinica*, pp. 361–386.

Lucas Jr, Robert E (1978). "On the size distribution of business firms". In: *The Bell Journal of Economics*, pp. 508–523.

Lundberg, Scott M et al. (2020). "From local explanations to global understanding with explainable AI for trees". In: *Nature machine intelligence* 2.1, pp. 56–67.

MacKinnon, James G and Halbert White (1985). "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties". In: *Journal of econometrics* 29.3, pp. 305–325.

Makridakis, Spyros, Steven C Wheelwright, and Rob J Hyndman (2008). *Forecasting methods and applications*. John wiley & sons.

Manjón-Antolín, Miguel C and Josep-Maria Arauzo-Carod (2008). "Firm survival: methods and evidence". In: *Empirica* 35, pp. 1–24.

Mata, Jose and Pedro Portugal (1994). "Life duration of new firms". In: *The journal of industrial economics*, pp. 227–245.

Mata, José and Pedro Portugal (2002). "The survival of new domestic and foreign-owned firms". In: *Strategic management journal* 23.4, pp. 323–343.

Mata, José, Pedro Portugal, and Paulo Guimaraes (1995). "The survival of new plants: Start-up conditions and post-entry evolution". In: *International Journal of Industrial Organization* 13.4, pp. 459–481.

McGowan, Müge Adalet, Dan Andrews, and Valentine Millot (2018). "The walking dead? Zombie firms and productivity performance in OECD countries". In: *Economic Policy* 33.96, pp. 685–736.

Merton, Robert C (1974). "On the pricing of corporate debt: The risk structure of interest rates". In: *The Journal of finance* 29.2, pp. 449–470.

Miller, Tim (2019). "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267, pp. 1–38.

Mitchell, Will (1994). "The dynamics of evolving markets: The effects of business sales and age on dissolutions and divestitures". In: *Administrative Science Quarterly*, pp. 575–602.

Modigliani, Franco and Merton H. Miller (1958). "The Cost of Capital, Corporation Finance and the Theory of Investment". In: *The American Economic Review* 48.3, pp. 261–297.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Moscatelli, Mirko et al. (2019). *Corporate default forecasting with machine learning*. Tech. rep. Bank of Italy, Economic Research and International Relations Area.

Mueller, Andreas I and Johannes Spinnewijn (2023). *The Nature of Long-Term Unemployment: Predictability, Heterogeneity and Selection*. Tech. rep. National Bureau of Economic Research. DOI: https://doi.org/10.3386/w30979.

Mullainathan, Sendhil and Jann Spiess (2017). "Machine learning: an applied econometric approach". In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Murray, Jared S (2017). "Log-linear Bayesian additive regression trees for categorical and count responses". In: *arXiv preprint, arXiv:1701.01503*.

Nickell, Stephen and Daphne Nicolitsas (1999). "How does financial pressure affect firms?" In: *European Economic Review* 43.8, pp. 1435–1456.

O'Neill, Eoghan (2024). "Type I Tobit Bayesian Additive Regression Trees for censored outcome regression". In: *Statistics and Computing* 34.4, pp. 1–19.

Odén, Anders, Hans Wedel, et al. (1975). "Arguments for Fisher's permutation test". In: *The Annals of Statistics* 3.2, pp. 518–520.

Ohlson, James A (1980). "Financial ratios and the probabilistic prediction of bankruptcy". In: *Journal of accounting research*, pp. 109–131.

Orbis (2020). *Orbis Company Information Across the Globe*. Bureau Van Dijk. A Moody's Analytics Company. URL: https://orbis.bvdinfo.com/.

Peek, Joe and Eric S Rosengren (2005). "Unnatural selection: Perverse incentives and the misallocation of credit in Japan". In: *American Economic Review* 95.4, pp. 1144–1166.

Penrose, E. (1959). *The Theory of the Growth of the Firm*. Oxford: Blackwell.

Prokhorenkova, Liudmila et al. (2018). "CatBoost: unbiased boosting with categorical features". In: *Advances in neural information processing systems* 31.

Rajan, Raghuram G and Luigi Zingales (Dec. 1995). " What Do We Know about Capital Structure? Some Evidence from International Data". In: *Journal of Finance* 50.5, pp. 1421–1460.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386*.

Riccaboni, Massimo, Xu Wang, and Zhen Zhu (2021). "Firm performance in networks: The interplay between firm centrality and corporate group size". In: *Journal of Business Research* 129, pp. 641–653.

Rozemberczki, Benedek et al. (2022). "The Shapley Value in Machine Learning". In: *arXiv preprint arXiv:2202.05594*.

Rungi, Armando and Francesco Biancalani (2019). "Heterogeneous Firms and the North–South Divide in Italy". In: *Italian Economic Journal*, pp. 1–23.

Saito, Takaya and Marc Rehmsmeier (2015). "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets". In: *PloS one* 10.3.

Schivardi, Fabiano and Guido Romano (2020). *A simple method to estimate firms liquidity needs during the COVID-19 crisis with an application to Italy*. CEPR Covid Economics: Vetted and Real-Time Paper.

Schivardi, Fabiano, Enrico Sette, and Guido Tabellini (2020). "Identifying the real effects of zombie lending". In: *The Review of Corporate Finance Studies* 9.3, pp. 569–592.

— (2021). “Credit Misallocation During the European Financial Crisis Credit Misallocation During the Crisis”. In: *The Economic Journal*.

Sharma, Anurag and Idalene F Kesner (1996). “Diversifying entry: Some ex ante explanations for postentry survival and growth”. In: *Academy of Management Journal* 39.3, pp. 635–677.

Shmueli, Galit et al. (2010). “To explain or to predict?” In: *Statistical science* 25.3, pp. 289–310.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje (2017). “Learning important features through propagating activation differences”. In: *International Conference on Machine Learning*. PMLR, pp. 3145–3153.

Soto-Simeone, Aracely, Charlotta Sirén, and Torben Antretter (2020). “New venture survival: A review and extension”. In: *International Journal of Management Reviews* 22.4, pp. 378–407.

Stam, Erik and Karl Wennberg (2009). “The roles of R&D in new firm growth”. In: *Small business economics* 33, pp. 77–89.

Strumbelj, Erik and Igor Kononenko (2010). “An efficient explanation of individual classifications using game theory”. In: *The Journal of Machine Learning Research* 11, pp. 1–18.

Sun, Jie and Hui Li (2011). “Dynamic financial distress prediction using instance selection for the disposal of concept drift”. In: *Expert Systems with Applications* 38.3, pp. 2566–2576.

Sutton, J. (1997). “Gibrat’s Legacy”. In: *Journal of Economic Literature* 35.1, pp. 40–59.

Teece, David J, Gary Pisano, and Amy Shuen (1997). “Dynamic capabilities and strategic management”. In: *Strategic management journal* 18.7, pp. 509–533. DOI: https://doi.org/10.1002/(SICI)1097-0266(199708)18:7<509::AID-SMJ882>3.0.CO;2-Z.

Tsai, Chih-Fong, Yu-Feng Hsu, and David C Yen (2014). “A comparative study of classifier ensembles for bankruptcy prediction”. In: *Applied Soft Computing* 24, pp. 977–984.

Tsai, Chih-Fong and Jhen-Wei Wu (2008). "Using neural network ensembles for bankruptcy prediction and credit scoring". In: *Expert systems with applications* 34.4, pp. 2639–2649.

Twala, BETH, MC Jones, and David J Hand (2008). "Good methods for coping with missing data in decision trees". In: *Pattern Recognition Letters* 29.7, pp. 950–956.

Udo, Godwin (1993). "Neural network performance on the bankruptcy classification problem". In: *Computers & industrial engineering* 25.1-4, pp. 377–380.

Ugur, Mehmet and Marco Vivarelli (2021). "Innovation, firm survival and productivity: the state of the art". In: *Economics of Innovation and New Technology* 30.5, pp. 433–467.

Unger, Jens M et al. (2011). "Human capital and entrepreneurial success: A meta-analytical review". In: *Journal of business venturing* 26.3, pp. 341–358.

Van der Laan, Mark J, Eric C Polley, and Alan E Hubbard (2007). "Super learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1.

Van Rijsbergen, Cornelis Joost (1979). *Information retrieval*. London: Butterworths.

Wang, Gang, Jian Ma, and Shanlin Yang (2014). "An improved boosting based on feature selection for corporate bankruptcy prediction". In: *Expert Systems with Applications* 41.5, pp. 2353–2361.

White, T Kirk, Jerome P Reiter, and Amil Petrin (2018). "Imputation in US manufacturing data and its implications for productivity dispersion". In: *Review of Economics and Statistics* 100.3, pp. 502–509.

Zhao, Y Lisa, Michael Song, and Gregory L Storm (2013). "Founding team capabilities and new venture performance: The mediating role of strategic positional advantages". In: *Entrepreneurship Theory and Practice* 37.4, pp. 789–814.