Big Data
Volume 9, Number 2, 2021
© Mary Ann Liebert, Inc.
DOI: 10.1089/big.2020.0047

ORIGINAL ARTICLE

Learning from Failure: Big Data Analysis for Detecting the Patterns of Failure in Innovative Startups

Maddalena Cavicchioli[1,*] and Ulpiana Kocollari[2]

Abstract

This article aims at identifying appropriate models for analyzing large datasets to serve a twofold goal: first, to better understand the dynamics impacting innovative startups' performance and their managerial practice and, second, to detect their patterns of failure. Therefore, we investigate the interaction of economic–financial, context, and governance dimensions of 4185 Italian innovative startups created from 2012 to 2015. Once startups have been grouped, we focus only on those that are unsuccessful. Then, failure patterns have been uncovered, integrating the use of factor and cluster analysis, where factor scores for each firm are used to identify a set of homogeneous groups based on clustering methods. The integrated use of those large-dimensional data techniques permits to classify items in rigorous ways and to unfold structures of the data, which are not apparent in the beginning. The analysis suggests that each pattern of failure is a multidimensional construct and, as a consequence can generate different managerial implications. Therefore, an effective handling of failure requires management to use appropriate interventions targeted at the challenges faced by that particular pattern of failure in the age of different firms.

Keywords: big data techniques; cluster analysis; factor analysis; failure patterns; innovative startups

Big Data in Management

A review of the recent literature

Although there is no unanimous definition of big data, there is a widely accepted awareness about the need to distinguish big data from what is commonly intended to be a large database. To the concept of big data are usually related three Vs, namely volume, variety, and velocity, which were introduced to better define and understand this notion.[1–3] Their analysis has become a promising practice in business involving a combination of diverse datasets and advanced analytic techniques that play an important role in influencing many aspects of business activities. With larger categories of data that can be collected and interpreted, companies are able to identify and answer better to markets' and stakeholders' requests. Further, the use of new techniques with higher performance and flexibility can meet the demand of more effective and efficient decision-making processes. In fact, the application of big data in organizations is influencing managerial practices and processes that are changing decision-making strategy, firms' culture, leadership, human resource management, and other management practices.[4]

According to the review of the literature carried out by Sheng et al.,[3] which examines the use of big data in business and management, eight main areas can be identified: general management, information management, marketing, operation research and management science, organization, industry study, and public sector study. The authors report that most of the studies apply the big data-driven approach in marketing activities and operational practices. On the contrary, only a few articles consider big data from a wider managerial perspective that investigates the strategic importance of their predictive use.

[1] *Department of Economics "Marco Biagi," University of Modena and Reggio Emilia & ReCent, Modena, Italy.*
[2] *Department of Economics "Marco Biagi," University of Modena and Reggio Emilia & Softech-ICT, Modena, Italy.*

*Address correspondence to: *Maddalena Cavicchioli, Department of Economics "Marco Biagi," University of Modena and Reggio Emilia & ReCent, Viale Berengario 51, Modena 41121, Italy,* E-mail: maddalena.cavicchioli@unimore.it

All the analyzed studies[3,5] reveal a growing interest and still great needs and space for big data research in the management sphere. Therefore, there are many opportunities and benefits that need to be explored for big data analysis in management development and business improvement.

In the light of these considerations, we argue that better predictions can be rooted in the big data elaboration instead of entrepreneurship instinct and experience, so that the data-driven approach can improve business performance.[6] This is even more true in the case of new ventures.[7]

Methods for analyzing big data

In the abstract of his 2001 paper in Statistical Science,[8] the statistician Leo Breiman writes about the difference between model-based and algorithmic approaches to statistics: "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown" (p. 199).

However, nowadays it is largely accepted that machine learning (ML), to which Breiman refers to as the algorithm modeling culture, may stand alongside with more traditional statistical methods.[9,10] Although the adoption of these methods in economics and management has been slower, they are now beginning to be widely used in empirical work and are the topic of a rapidly increasing methodological literature. Although relevant more generally, the methods developed in the ML literature have been particularly successful in big data settings, where we observe information on a large number of units, many pieces of information on each unit, or both, and often outside the simple setting with a single cross-section of units.

Even though there are cases where using simple off-the-shelf algorithms from the ML literature can be effective, there are also many cases where this is not the case. The ML techniques often require careful tuning and adaptation to effectively address the specific problems that economists are interested in. Perhaps the most important type of adaptation is to exploit the structure of the problems. Statistics and econometrics have traditionally placed much emphasis on these structures and developed insights to exploit them with traditional structural modeling, whereas ML has often placed little emphasis on them. However, a combination of the two approaches is what is needed to unfold complicated structures from a large multidimensional set of data.

The Case of Innovative Startups: The Importance of Combining Economic and External Context Data

Italian innovative startups

This study examines the set of startups in Italy. Over the past decades, this country has witnessed first the introduction and then the expansion of innovative business models for new ventures, in a variety of forms, as an integral part of its economic context. The scenario has been further reinforced by the creation of new legal forms of organizations that are absolute firsts on the international economic stage. In the sequel, we provide a brief overview of the new form of enterprise introduced during the past few years, to highlight the opportunities and needs that it implies.

One very interesting category of startup examined in this study is that of innovative ones. The Italian Decree Law 179/2012 introduced a new body of legislation governing the foundation and growth of innovative startups. Specifically, article 25, subsection 2 defines an innovative startup as a capital enterprise, which may also be a cooperative, incorporated under Italian law, shares or stakes that are not listed on a regulated market or on a multilateral trading system, which meets specific requirements. Within this broader definition of an innovative startup, the law identifies two additional types of enterprise that are awarded greater fiscal benefits for their potential investors: startups with social goals and high-tech startups in the energy industry.

We are, therefore, looking at a type of enterprise that has distinct characteristics from its foundation and is then in need of a measurement process capable of identifying, calculating, and monitoring its potential. This measurement process should integrate the company's economic performance and its context variables.

From an operational point of view, an innovative startup must meet all the criteria for a startup and also fulfil an additional condition related to its innovation in its area of business, to be qualified as such. In particular, for innovative startups it is mandatory to comply with the "cumulative" prerequisites (meaning that they must all be fulfilled) listed in the aforesaid article 25, subsection 2, and the "alternative" prerequisites at the same point. Those cumulative prerequisites are summarized next:

- The company must have been incorporated, and have been engaging in the business concerned, for no more than 60 months, meaning 5 years;
- it must have its registered office and operations center in Italy;

- from the innovative startup's second year in business, the total value of annual production must not exceed 5 million euros;
- it must not distribute or have distributed profits;
- its exclusive or prevalent corporate purpose must be the development, production, and sale of innovative, high-tech products or service;
- it must not have been incorporated from a merger or corporate break-up, or further to the transfer of ownership of a company or company division;
- it must operate in the sectors envisaged by the law. Moving on to the second category of requirements (which are alternatives), the law requires the company to fulfil at least one of the following conditions:
- It must spend at least 15% of the cost or total value of its production (whichever is greater) on research and development;
- at least one-third of the total workforce must consist of highly qualified people, who may be formal employees or freelance associates of any kind;
- it must be the owner, registered holder, or licensee of at least one patent.

These characteristics imply specific expedients for analyzing and evaluating the activities of these firms. In particular, a set of qualitative context variables are added to the traditional economic and financial criteria to identify their innovative activities.

Investigating the innovative startups' performance and failure

Measuring the performance of startups is an important task, as they can be a source of stable job creation. Further, startups' improvement in performance is critical to their survival and growth.[11] The dimensions that are most frequently used by researchers, such as annual revenues, growth in sales, and number of employees alone are not enough to explain and detect startups' development paths. According to Chorev and Andersin,[12] the success of a firm can often be influenced by external factors such as the government support,[13] competitive rivalry, innovation, and industry.[14] The external forces, combined with the economic internal ones, may act as a driver behind the growth or failure of a new venture. In fact, a growing body of research has focused on the external environmental conditions[15,16] as the elements that might condition the success or failure of a startup. In this light, we particularly consider in our study: the geographic location of the startup, as it can benefit a specific government's support that facilitates growth in that area[17] or can help firms to be close to strategic suppliers and customers[16]; the industry category since startups of the same industry have similar behaviors toward technological changes and toward growth.[14] A set of variables that are representative of ownership and/or corporate governance characteristics are also considered with regard to the managerial experience of the owners and the governance models that can influence performance.[18–21]

Unlike established firms, which have consolidated models for analyzing their level of viability and survival, new ventures are subject to a liability of newness where their survival is a critical issue.[22] In particular, understanding new business failure still represents a challenge for researchers since most studies deal with prediction of failure[23] and only a few focus on its understanding. Most of these prediction models depend on accurate quantitative data over several years before failure[24] that are problematic, especially for startups and small ventures that are known to be weak in financial data records. Further, the multidimensional nature of failure implies that numerous variables of a nonfinancial kind should be analyzed for marking the causes of failure.[25] In their investigation and analysis of bankrupt firms, Thornhill and Amit[26] compared industry change, general management, financial management, and market development variables associated with different stages of business development and suggest that young firms are more likely to suffer from resource and capability deficiencies than older firms, which is essence of the "liability of newness."[25,27]

A firm's failure generally does not stem from a single factor but it is the result from an accumulation of decisions, actions, and commitments that become knotted in self-perpetuating dynamics.[28] A precondition refers to conditions that must exist or be established before something can occur, thus it is a prerequisite. Francis and Desai[29] refer to preconditions as contextual factors. Therefore, the contextual factors as triggers of failure are specifically relevant for understanding decline. All these studies underline the relevance of external variables for the analysis of startups' development, warning that a combination of different factors can better explain why new ventures succeed or fail. From a methodological perspective, the need of a new framework rises, to combine both the two dimensions: economic/financial and context/external.

Motivated by those reasons, the aim of this article is to implement data-mining techniques to better understand the dynamics of economic, context, and governance dimensions and their influence over the innovative startups' performance and finally detect different patterns of failure. We focus on financial, governance, and context data of all 4185 Italian innovative startups created from 2012 to 2015. We first use the data-mining clustering technique to identify groups of startups that show similar behavior. Second, we focus only on those unsuccessful startups where failure patterns have been uncovered, integrating the use of factor and cluster analyses. Those techniques are, in fact, useful in reducing both the variables' dimension and the units' dimension. Particularly, factor analysis is able to synthetize a variety of collected aspects (the multivariate collection of data) into a few latent dimensions that are non-redundant expressions of the initial variables. Then, groups are formed by using cluster analyses, where the factor scores for each firm are used to identify a set of homogenous groups.

Analysis on Large Multidimensional Data

Methodology for data mining

Given the quantity and variety of variables, the need to reduce and synthesize their multidimensionality has been taken into account in the selection and construction of the methodology used for data processing and for their analysis. To detect failure patterns, two different statistical approaches are applied to identify different processes. Particularly, the analysis is based on the integrated use of factor and cluster analyses, where the factor scores for each firm are used to identify a set of homogenous groups based on cluster analysis. For factor analysis, the unweighted least-squares extraction method with varimax rotation is used and the number of factors is determined by using the eigenvalues exceeding one rule. For the subsequent cluster analysis, *K*-medians clustering is chosen by using the L2/Euclidian dissimilarity measure and initial group centers are selected as *K* well-spaced observations. Cluster analysis was conducted with different *K* values starting from 2 and evaluated with the R^2 criterion.

A major topic in the ML literature is unsupervised learning. In this stream of the literature enters k-means clustering,[30,31] whose goal is, given a set of observations on different aspects X_i, to partition the feature space into subspaces. Those subspaces should be obtained by having units that are homogeneous within the cluster and as much heterogenous between clusters.

Consider the case where we wish to partition the initial space into G subspaces or clusters. We start choosing centroids $c_1, \ldots, c_G$ and then assign units to the cluster based on their proximity to the centroids. The algorithm works as follows. We start with a set of G centroids, $c_1, \ldots, c_G$, elements of the initial space that are sufficiently separated over this space. Given a set of centroids, we assign each unit to the cluster that minimizes the distance between the unit and the centroid of the cluster:

$$C_i = \arg \min_{g \in \{1,\ldots,G\}} || X_i - c_g||^2.$$

Then, we update the centroids as the average of the X_i in each of the clusters:

$$c_g = \sum_{i:C_i=g} X_i \;/\; \sum_{i:C_i=g} 1.$$

We repeatedly iterate between the previous two steps. The choice of the number of clusters is not a straightforward task, because there is no direct cross-validation method to assess the performance of one value versus the other. It is common use to compare the value of the index R^2, computed as 1 minus the ratio between the variance within clusters over the total variance, adjusted by degrees of freedom.

Data and results

As a first step, we analyzed the context variables of all the innovative startups active in Italy at the date of the investigation. We consider the extensive population of 4185 innovative startups in Italy in 2015, using firm-level data obtained from *Aida Bureau van Dijk* database. The set of variables used to classify items are categories of geographical areas, industry, number of startups' shareholders, and number of startups' shareholders that are also managers. Their characterization in terms of geographical areas and industry is reported in Figure 1.

With regards to the number of shareholders, the frequency distributions is 0: 7.5%, 1–2: 37.2%, 3–7: 45.5%, >7: 9.7%. Finally, in terms of shareholders that are also managers, the distribution is 0: 52.8%, 1: 27.2%, 2: 18.2%, >2: 1.8%.

Based on those four context variables, items are classified by using *K*-means methodology, as presented in the previous subsection. To identify the number of clusters, we make use of the R^2 index, computed as 1 minus the ratio between the variance within clusters over the total variance, adjusted by degrees of freedom. For a number of clusters G equal to 2, the R^2 is equal to

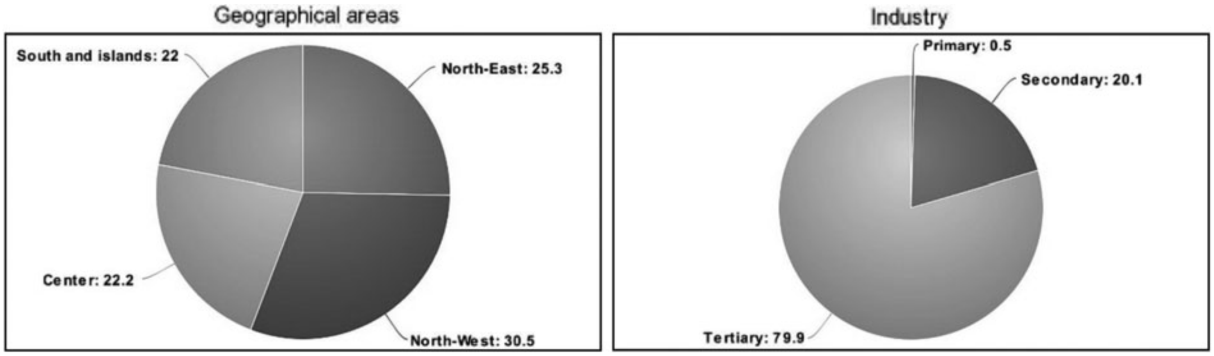

FIG. 1. Distributions of the cases with respect to geographical areas (left) and industry (right).

0.7938; whereas for $G=3$, the index jumps at 0.9905, which indicates an almost perfect partition. So, we stay at $G=3$ clusters and convergence is reached after five iterations. The first identified cluster is a subset of 1222 startups mainly located in the center and south area (88%), belonging to the tertiary industry for the 83%, with the lowest number of managers (zero at the 86%) and a number of shareholders greatly above the mean value. With regards to the economic features, they are characterized by low revenues and at the same time low indebtedness.

The second cluster is the most numerous (2152 U), with all of them located in the north of Italy and active in the tertiary (77%) and secondary (22%) sectors. They are characterized by having up to one single manager (79%) and a number of shareholders greatly above the mean value. In particular, they show higher revenues with respect to the other two clusters and an average level of indebtedness.

The third group is composed by 811 startups, mainly in the tertiary sector (81%) and present in the center-south of Italy (96%). Differently from the first cluster, they are characterized by having a low number of shareholders and the higher number of managers. They have obtained good revenues (but less than the second cluster) with an average level of indebtedness.

With particular attention to the failed startups in the three groups, we observe that in the first cluster there are no failed startups, in the second there are only 4% failed ones, whereas in the third the number of

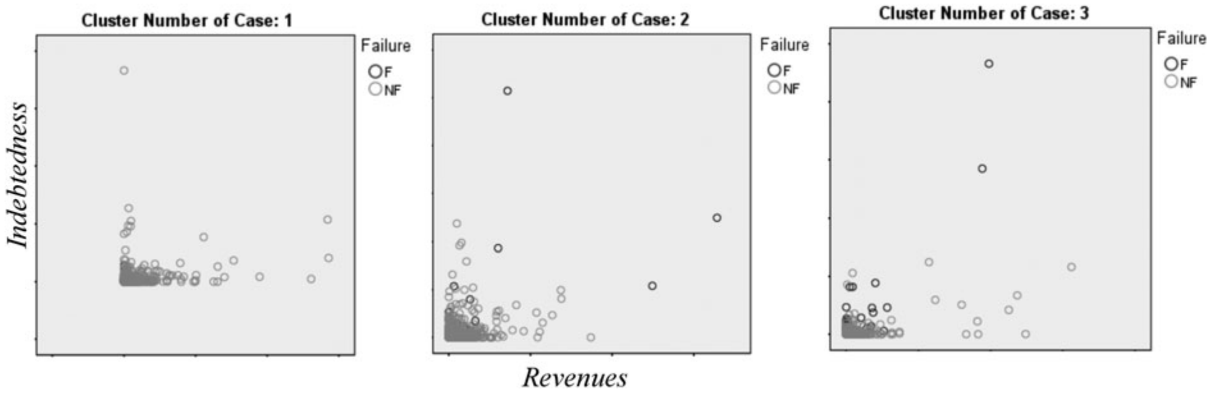

FIG. 2. Scatterplots of identified clusters with respect to two economic dimensions (revenues on the *x*-axis and indebtedness on the *y*-axis), highlighting the failed startups (black dots) and the nonfailed ones (grey dots).

Table 1. Tag and description of the variables in the study

Variable	*Description*
NIA	Net income to total assets
EBITA	Earnings before interest, taxes, amortization, and depreciation
DA	Total debt to total assets ratio
SA	Turnover/total assets
CACL	Current assets/current liabilities × 100
CCL	Cash and equivalents/current liabilities × 100
NWCA	(Current assets − current liabilities)/total assets × 100
EBITI	Earnings before interest and taxes/financial expenses
Nr. SRH	Total number of shareholders
Nr. EMP	Total number of employees
Nr. Mng	Number of shareholders that are managers
INDUSTRY	1-Manifacturing; 2-Services; 3-ICT; 4-Commerce; 5-Agriculture
AREA	Geographical area in which the startup operates

Note that industry's categories are further specified into five classes.

failed startups reaches 14%. This is shown graphically in Figure 2, which also reports the two economic dimensions: levels of revenue and indebtedness.

There results show three distinct clusters and two main categories of failure. However, although it is undeniable that external forces are relevant for the analysis of startups' paths of development, these factors alone are not sufficient to explain the processes standing behind new ventures' survival or failure. Further, deep investigation combining these context variables with further economic/financial indicators is needed.

Detecting the Patterns of Failure in Innovative Startups

After the detection of the failure main cluster and its main characteristics, in this section, we analyze the failed startups considering both their economic and context variables to identify multifaceted patterns of failure. When only bankrupted startups are included in the analysis (particularly those founded in 2012 and bankrupted at the latest in 2015, which are actually present in clusters 2 and 3), the considered variables are reported and described in Table 1.

Overall, 23.2% of the failed startups are located in the North-East area, 35% in the North-West area, 23.2% in the center of Italy, and 18.6% in the South (including the two main islands of Sicily and Sardinia). With regard to industry, 36% of them belong to the service industry, 44.3% to the Information and Communications Technology (ICT) industry, 13.8% to the manufacturing sector, 4.9% are trading firms, and the remaining 1% belong to the agricultural sector.

The extracted failure processes based on the consecutive application of factor and cluster analysis are reported in Table 2. The first things to notice are the increase in the number of factors with the firm's age as well as the proportion of variance explained by factors; second, the decrease of the number of failure processes as the firm's age increases. In fact, the number of different patterns of failure reduces from three to two with the increase of the firm's age.

Moreover, important evidence was found about the presence of different patterns of failure in different sectors for most firms' age groups, whereas the patterns of failure are not generally associated with the geographical area in which the startups are born. The numerosity of shareholders features different patterns of failure only in mature startups.

For all extracted failure processes, the respective median values for the variables in the study are examined for different lifetimes and reported in Tables 3–6 next.

Failure processes for 1-year-old firms are clearly distinct in the two most numerous clusters in which the "symptoms" clearly symbolize acute failure. In particular, the firms belonging to one main cluster are characterized by low performance (net income to total assets [NIA]: −24.6% and earnings before interest, taxes, amortisation and depreciation [EBITA]: −23.6%), whereas the others in the second main cluster are characterized by poor financial structure (current assets/current liabilities × 100 [CACL]: −26.2% and cash and equivalents/current liabilities × 100 [CCL]: −21%). These distinctive characteristics are also reported in the management structure of

Table 2. Extracted failure processes

Firm age (years)	*No. of factors and variance explained (%)*	*Bartlett's significance*	*Firms in cluster (share)*			R^2	*Total firms in the analysis*
			1	*1*	*3*		
1	3; 76.8	0.000	28 (50%)	1 (2%)	27 (48%)	0.74	56
2	6; 83.8	0.000	3 (5%)	50 (79%)	10 (16%)	0.86	63
3	8; 88.6	0.000	9 (24%)	28 (76%)		0.63	37
4	9; 90.7	0.000	5 (21%)	19 (79%)		0.71	24

Table 3. Median of variables in different failure processes for 1-year-old firms

Variable/failure process (share)	*1.1 (50%)*	*1.2 (2%)*	*1.3 (48%)*
NIA	21.49%	61.95%	−24.58%
EBITA	20.59%	60.46%	−23.59%
DA	0.86	1.31	0.84
SA	0.21	0.14	0.22
CACL	−26.23%	707.88%	0.99%
CCL	−20.85%	701.99%	−4.37%
NWCA	−61.15%	242.22%	54.45%
EBITI	0.07	0.90	−0.10

CACL, current assets/current liabilities × 100; CCL, cash and equivalents/current liabilities × 100; DA, total debt to total assets ratio; EBITA, earnings before interest, taxes, amortisation and depreciation; EBITI, earnings before interest and taxes/financial expenses; NWCA, (Current assets − current liabilities)/total assets × 100; ML, machine learning; NIA, net income to total assets; SA, turnover/total assets.

the two main clusters. In fact, in the first main cluster the shareholders, who are also managers, are less than in the majority of firms in the second main cluster.

The failure processes detected for 2-year-old firms, similarly to the previous group, are three.

Considering the industries in which the firms of each cluster operate, a distinction can be made: The firms of the first cluster are all manufacturing ones, those of the third cluster are mainly (60%) ICT innovative ventures, and the second cluster is a mixed one. The failed manufacturing firms of the first cluster are characterized by negative profitability (NIA second year: −2.83), very low liquidity [CCL of the second year: −53.7%; (Current assets − current liabilities)/total assets × 100 (NWCA) of the second year: −243%], and unsustainable financial structure (total debt to total

Table 4. Median of variables in different failure processes for 2-year-old firms

Variable/failure process (share)	*2.1 (5%)*	*2.2 (79%)*	*2.3 (16%)*
NIA1	20.32%	−8.66%	37.22%
NIA2	−283.44%	9.64%	36.83%
EBITA1	16.00%	−8.54%	37.90%
EBITA2	−284.30%	9.28%	38.90%
DA1	0.90	0.04	0.09
DA2	2.38	0.18	0.17
SA1	0.63	0.12	0.79
SA2	0.63	0.13	0.82
CACL1	−23.35%	1.21%	0.98%
CACL2	−48.17%	−23.23%	130.61%
CCL1	−42.70%	4.43%	−9.33%
CCL2	−53.72%	−22.32%	127.70%
NWCA1	−198.90%	5.71%	31.11%
NWCA2	−242.57%	4.29%	51.31%
EBITI1	−2.10	0.10	0.14
EBITI2	−2.23	0.07	0.32

Table 5. Median of variables in different failure processes for 3-year-old firms

Variable/failure process (share)	*3.1 (24%)*	*3.2 (76%)*
NIA1	77.06%	−24.77%
NIA2	7.39%	−2.38%
NIA2	9.57%	−3.07%
EBITA1	71.86%	−23.10%
EBITA2	27.01%	−8.68%
EBITA3	−6.93%	2.23%
DA1	0.02	0.00
DA2	0.70	0.22
DA3	0.63	0.20
SA1	0.26	0.08
SA2	0.40	0.13
SA3	0.20	0.06
CACL1	190.45%	−61.22%
CACL2	155.25%	−49.90%
CACL3	274.56%	−88.25%
CCL1	93.71%	−30.12%
CCL2	116.00%	−37.29%
CCL3	268.96%	−86.45%
NWCA1	27.74%	−8.92%
NWCA2	83.87%	−26.96%
NWCA3	110.04%	−35.37%
EBITI1	−0.16	0.05
EBITI2	0.44	−0.14
EBITI3	0.22	−0.07

Table 6. Median of variables in different failure processes for 4-year-old firms

Variable/failure process (share)	*4.1 (21%)*	*4.2 (79%)*
NIA1	22.15%	−5.83%
NIA2	12.86%	−3.38%
NIA2	14.07%	−3.70%
NIA4	−59.43%	15.64%
EBITA1	−71.21%	18.74%
EBITA2	21.29%	−5.60%
EBITA3	17.06%	−4.49%
EBITA4	−49.83%	13.11%
DA1	0.25	0.06
DA2	0.15	0.04
DA3	0.25	0.07
DA4	0.09	0.02
SA1	0.41	0.11
SA2	0.20	0.05
SA3	0.15	0.04
SA4	0.22	0.06
CACL1	4.25%	−1.12%
CACL2	62.51%	−16.45%
CACL3	23.05%	−6.06%
CACL4	74.64%	−19.64%
CCL1	−8.53%	2.25%
CCL2	37.05%	−9.75%
CCL3	16.90%	−4.45%
CCL4	109.87%	−28.91%
NWCA1	28.05%	−7.38%
NWCA2	27.64%	−7.27%
NWCA3	67.36%	−17.73%
NWCA4	90.26%	−23.75%
EBITI1	−0.76	0.20
EBITI2	0.57	−0.15
EBITI3	0.26	−0.07
EBITI4	0.04	−0.01

assets ratio [DA] of the second year: 2.38) 1 year before failure. The third group of firms (mainly ICT) shows a pattern of slight growth (ΔEBITA: 1%) but remains unsustainable in terms of financial structure (ΔCACL: 128%, ΔCCL: 118%). Finally, the second group is different in the sharp change when passing from the first to the second year both in performance and in financial structure that is not sufficient to sustain the scaling of the activity.

For 3-year-old firms, only two patterns were detected.

The first cluster figures out a clear poor performance after the first year of activity (ΔNIA of the second year: 70%, ΔEBITA of the second year: 22%). Financial ratios show a break indicated by the sharp decline of the ratios 2 years before failure (ΔCACL of the second year: −45%, ΔDA of the second year: 0.7). This trend is also registered in the second cluster but the first pattern differs from the second one mostly with respect to the entity of the ratios' difference in the last years of the activity (ΔCACL of the third year: −35%, ΔDA of the second year: 0.20, and ΔDA of the third year: 0.02). This difference can be attributed to the characteristics of the governance structure. In fact, in the second cluster we observe the presence of more management competencies through the owners of the firms, whereas the first cluster counts a maximum of one manager that is also a shareholder of the startup.

Four-year-old firms also witness two different failure processes.

Both groups of firms have a remarkably lower productivity of assets [mean of turnover/total assets (SA) over the 4 years: 0.25 − first cluster, mean of SA over the 4 years: 0.07 − second cluster]. In the first pattern, a gradual decline is shown in case of financial ratios, which worsen and attain poor values 1 year before failure (ΔCACL of the last year: 49%, ΔDA of the last year: −0.14). In case of the second and more numerous process, the downturn starts much earlier and many financial ratios have low values throughout the startups' entire life cycle (mean of CACL over the 4 years: −11% mean of DA over the 4 years: 0.05). Note that one characteristic of the more long-lived firms is to be located in the North (67%) with respect to any other areas. This fact underlines the importance of the context that accentuates the existence of an ecosystem that supports entrepreneurial projects.

Conclusions

We apply data-mining techniques to uncover groups and patterns of failure in the context of Italian innovative startups. The patterns of firms' performance identified with our analysis favor the use of different dimensions characterizing startups' life, such as economic–financial indicators and context data. We are able to conclude that it is reasonable to argue that understanding failure for startups has utility for detecting patterns, which differ in terms of economic, context, and governance dimensions. Of course, there will be common features that transcend the various patterns of failure, as the startup phases of business will inevitably carry with them vestiges of the business idea and the financial structure. Some features such as financing issues are relatively persistent, despite the changing needs as startups move from one stage to another. However, other features are unique to each pattern since the main issues in which the startup's decision-making process takes place will change significantly over the different stages of business idea development. These elements, if contextualized, can pinpoint the presence of facts or conditions that mark the possible causes of failure and they can be referred to as "warning signs," which are not evident if considered detached from the process in which they occur. Even if the detected features cannot predict failure *per se*, they might be considered as purely indicators that failure causes are present within a certain context. With the patterns we identified, we can provide some more specific recommendations for the industry to predict the startups' failure. For example, using the patterns represented in the Cluster 4.1 and 4.2 we can individuate the following "warning signs."

The failure pattern outlined in Cluster 4.1 appears evident: The destruction of wealth, resulting from negative profitability, is not mitigated by the constant reduction in debt; short-term liabilities have been used to finance fixed investments with a consequent situation of financial liquidity difficulties. In other words, the innovative startups of Cluster 4.1 have tried to restore their equilibrium, after the decrease in profitability, only through the increase in current liabilities, instead of acting also on business costs, in a structural way. This was accompanied by an increase in the incidence of current assets on total net assets, which resulted in an increase in suspended costs (inventories) and noncollections (commercial credits).

The pattern of failure of the innovative companies in Cluster 4.2 is primarily influenced, unlike Cluster 4.1, by the type of governance adopted; in particular, these firms have a greater number of managers and employees. Profitability is positive and innovative startups have a good margin on sales, but the situation regarding the turnover of invested capital is critical; in fact, in terms of speed if disinvestment of a firm's resources is slow, that is, innovative startups are not able to exploit the resources invested and, in particular, their production capacity. This is combined with the increase in debt year after year, since there is a need to remedy a volume of activity that is not appropriate to the structure (oversizing) and it should be noted that the increase in the use of debt capital is also due to the management of net working capital, which shows a short-term asset that does not cover short-term liabilities (as a result, short-term payments have been met by a reduction in long-term debt).

Summing up the study identifies the spread of two phenomena, which presents the Italian startup ecosystem: "dwarfism" and the presence of so-called "zombie startups." Dwarfism represents the situation in which startups do not fully develop their growth potential, that is, they do not enter the expansion phase; this is the case of subcontracting companies, which have become specialized cooperators and suppliers of innovative components. On the contrary, zombie startups are innovative startups that survive beyond 3 years with minimal turnover and activity, compared with the legislative framework. Their presence is reflected on the scarce investments in research and innovation, since they are not productive and absorb capital.

Therefore, by studying the patterns of failure as a mix of profitability, context, and governance dynamics over the different age of failure, this work adds a complementary perspective to the current state of knowledge on startups' failure. The possibility to consider at the same time various aspects of a phenomenon becomes an advantage if opportune techniques are put in place. In fact, to detect and monitor early warning signals that could lead to failure, entrepreneurs should have clear the main patterns of failure that their business is mostly exposed to. This could be done having in mind specific models for analyzing a large amount of data that are able to synthetize the multidimensionality of the phenomenon into a few determinants of the business and its stage.

Acknowledgments

The authors would like to thank participants at the International Conference on Big Data in Business (ICBIB 2019) held in London (United Kingdom), October 29–31, 2019, for their useful comments and remarks. The authors would like to thank the Editor-in-Chief of the journal, Professor Zoran Obradovic, and the anonymous reviewers for their valuable comments and suggestions, which improved the final version of the article.

Author Disclosure Statement

No competing financial interests exist.

Funding Information

Maddalena Cavicchioli acknowledges financial support by "Fondo di Ateneo per la Ricerca" FAR (2019) research grant of the University of Modena and Reggio Emilia, Italy.

References

1. Kwon O, Lee N, Shin B. Data quality management, data usage experience and acquisition intention of big data analytics. Int J Inf Manage. 2014; 34:387–394.
2. Gunasekaran A, Papadopoulo, T, Dubey R, et al. Big data and predictive analytics for supply chain and organizational performance. J Bus Res. 2017;70:308–317.
3. Sheng J, Amankwah-Amoah J, Wang X. A multidisciplinary perspective of big data in management research. Int J Prod Econ. 2017;191:97–112.
4. Davenport TH. Big data at work: Dispelling the myths, uncovering the opportunities. Boston, MA: HBS Press 2014.
5. Chen DQ, Preston DS, Swink M. How the use of big data analytics affects value creation in supply chain management. J Manage Inf Syst. 2015;32: 4–39.
6. McAfee A, Brynjolfsson E. Big data: The management revolution. Harvard Bus Rev. 2012;90:60–68.
7. Provost F, Webb GI, Bekkerman R, et al. A data scientist's guide to start-ups. Big Data. 2014;2:117–128.
8. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat Sci. 2001;16:199–231.
9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Berlin: Springer 2009.
10. Efron B, Hastie T. Computer age statistical inference, vol. 5. Cambridge, United Kingdom: Cambridge University Press 2016.
11. Brush CG, Vanderwerf PA. A comparison of methods and sources for obtaining estimates of new venture performance. J Bus Ventur. 1992;7: 157–170.
12. Chorev S, Anderson A. Success in Israeli high-tech start-ups: Critical factors and process. Technovation. 2006;26:162–174.
13. Arruda C, Silva V, Costa V. The Brazilian entrepreneurial ecosystem of startups: An analysis of entrepreneurship determinants in Brazil as seen from the OECD pillars. J Entrep Innov Manag. 2013;2:17–57.
14. Cowling M, Fryges H, Licht G, Murray G. Survival of new technology based firms in the UK and Germany. Front Entrep Res. 2006;26:1–11.
15. Timmons J, Spinelli S. New venture creation: Entrepreneurship for the 21st Century. New York: McGraw-Hill-Irwin 2004.
16. Hormiga E, Batista-Canino R, Sánchez-Medina A. The role of intellectual capital in the success of new ventures. Int Entrepreneurial Manag J. 2010;7:1–22.
17. Pugliese R, Bortoluzzi G, Zupic I. Putting process on track: Empirical research on start-ups? growth drivers. Manage Decis. 2016;54:1633–1648.
18. Chen HH. The timescale effects of corporate governance measure on predicting financial distress. Rev Pacific Basin Finan Markets Policies. 2008;11:35–46.
19. Deng X, Wang Z. Ownership structure and financial distress: Evidence from public-listed companies in China. Int J Manage. 2006; 23:486.
20. Fich EM, Slezak SL. Can corporate governance save distressed firms from bankruptcy? An empirical analysis. Rev Quant Fin Acc. 2008;30: 225–251.

21. Lee YJ, Roth WM. Making a scientist: Discursive "doing" of identity and self-presentation during research interviews. Qual Soc Res. 2004;5.
22. Gilbert BA, McDougall PP, Audretsch DB. New venture growth: A review and extension. J Manage. 2006;32:926–950.
23. Ooghe H, De Prijcker S. Failure processes and causes of company bankruptcy: A typology. Manage Dec. 2008;46:223–242.
24. Muller G, Steyn-Bruwer BW, Hamman WD. Predicting financial distress of companies listed on the JSE—A comparison of techniques. South Afr J Bus Manage. 2006;40:21–32.
25. Shepherd DA, Wiklund J, Haynie JM. Moving forward: Balancing the financial and emotional costs of business failure. J Bus Ventur. 2009;24: 134–148.
26. Thornhill S, Amit R. Learning about failure: Bankruptcy, firm age, and the resource-based view. Organ Sci. 2003;14:497–509.
27. Zacharakis AL, Meyer GD, DeCastro J. Differing perceptions of new venture failure: A matched exploratory study of venture capitalists and entrepreneurs. J Small Bus Manage. 1999;37:1.
28. Kanter RM. Leadership and the psychology of turnarounds. Harvard Bus Rev. 2003;81:58–69.
29. Francis JD, Desai AB. Situational and organizational determinants of turnaround. Manage Decis. 2005;43:1203–1224.
30. Gatignon H (ed). Multivariate Normal Distribution. In: Statistical Analysis of Management Data. Springer, Boston, MA, 2013, pp. 9–29.
31. Alpaydin E. Introduction to machine learning. Cambridge, MA: MIT Press 2009.

Cite this article as: Cavicchioli M, Kocollari U (2021) Learning from failure: Big data analysis for detecting the patterns of failure in innovative startups. *Big Data* 9:2, 79–88, DOI: 10.1089/big.2020.0047.

Abbreviations Used

CACL	=	current assets/current liabilities × 100
CCL	=	cash and equivalents/current liabilities × 100
DA	=	total debt to total assets ratio
EBITA	=	earnings before interest, taxes, amortisation and depreciation
ICT	=	Information and Communications Technology
ML	=	machine learning
NIA	=	net income to total assets
SA	=	turnover/total assets