Journal of the Knowledge Economy
https://doi.org/10.1007/s13132-024-02468-w

X-Techs: What Matters for the Survival of Brazilian Startups?

Kelly Lima Fonseca Gonçalves[1] · Berislav Andrlić[2] · Rodrigo Franco Gonçalves[1] · Irenilza de Alencar Nääs[1]

Received: 21 June 2023 / Accepted: 22 October 2024

Abstract

The factors contributing to the survival of startups are an emerging area of research, requiring a deeper understanding of the variables that influence their success. This study employs data mining techniques to analyze the relationship between business segments, target audiences, income models, business stage, and the survival rate of startups in Brazil. Starting from a public, unique Brazilian startups research database with 12,207 listed startups, we create a dataset of 2249 technological-based startups in the most representative business segments, such as education (Edutechs), finance (Fintechs), biology (Biotechs), and others. We call the business segment the "X" variable, the term's origin in X-Tech. Our initial hypothesis was that the business segment ("X" variable) was a determinant of business survival. This dataset was used to construct a random forest model using Rapidminer software to predict which independent variables are more relevant to the survival of startups. The findings reveal, with an accuracy of 70% and $\kappa = 0.72$, that the choice of target audience (primarily B2C and B2B) and income model (particularly the marketplace model) are more influential in determining the survival of Brazilian technological-based startups. The marketplace model, offering visibility, cost-effectiveness, and convenience, emerges as a crucial factor, especially with a B2C or B2B target audience. Both primary and secondary variables suggest that positioning a startup on a marketplace platform targeting a B2B or B2C audience is more likely to enhance its chances of survival in Brazil. The study also shows that the business segment, the "X" of the X-techs, was not relevant to the survival rate.

Keywords Start-up · Business survival · Digital economy · Business longevity

JEL Classification L26

Extended author information available on the last page of the article

Published online: 28 December 2024

Introduction

The year 2020 was forcibly the year of global digital transformation. The spread of the SARS-COV-2 virus has pushed humanity into adopting information technology tools, prompting a significant digital shift. Social isolation caused by the pandemic moved nearly all companies' offices to employees' and directors' homes in a home office movement. Meetings began to be held on videoconference platforms, and traditional schools hired educational platforms for distance learning. In this way, sales and purchases were made through websites or social networks with delivery, while families' leisure was limited to streaming channels (Higuchi & Maehara, 2021; Steinbach et al., 2021).

The swift evolution of consumer and corporate needs demands agile innovation, pushing businesses toward new market paradigms and disrupting established models. This growing uncertainty drives companies to undergo digital business transformation (DBT) to adapt their core strategies. Technology-driven innovation has become crucial for economic and social progress in the face of intense competition and rapidly changing consumer preferences. Many businesses have embraced digital transformation principles at various levels, from operational enhancements like developing new products or services to strategic overhauls of their entire business models (Matt et al., 2015).

In this change process, digital transformation is critical in improving the quality and diversity of new products and services and strengthening relationships between companies, consumers, and the government. Innovating and generating new products is critical for organizational survival in the current business landscape. This has spurred the rise of startups, nascent ventures that aim to commercialize innovative ideas. Often focused on technology, these startups navigate uncertain markets (Barbosa & Ramos, 2021; Donda, 2020). To succeed, they must possess technical and business expertise, strategically allocating resources to create products or services with market appeal.

Digital transformation and the rise of startups are prominent phenomena in the contemporary business landscape, representing more than mere technological tools. These concepts encompass a holistic shift influencing individuals, organizations, behaviors, operations, management practices, and hierarchical structures. Embracing digital transformation can confer a competitive advantage, enabling firms to enhance economic performance indicators (Westerman et al., 2012; Hess et al., 2016; Boneva, 2018; Heavin & Power, 2018), adapt to evolving consumer trends (Kim et al., 2017; Ismail et al., 2016; Dremel et al., 2017; Von Leipzig et al., 2017); and even redefine competitive boundaries (Schwertner, 2017). Therefore, startups have a clear advantage because they do not have legacy or corporate restrictions. Innovation and adopting new technologies are fundamental aspects of these new businesses (Hatada, 2021). Consequently, technology-driven startups in different business segments, such as education (Edutechs), finance (Fintech), insurance (Insurtechs), and biology (Biotechs), have rapidly transformed the landscapes of their respective industries.

In light of the accelerated digital transformation permeating various facets of society, coupled with the intrinsic nature of startups to seek valuable propositions and build scalable, replicable business models (Mercandetti et al., 2017), technology-driven startups are gaining increasing prominence and are often referred to as "Tech companies." The fusion of "financial" and "technology" gave rise to the widely recognized term "Fintech," representing startups that strive to revolutionize and enhance services within the financial system. These companies operate with significantly reduced costs compared to conventional financial institutions. The insurance industry, propelled by the Fintech wave, has witnessed rapid expansion, with the emergence of companies offering innovative "Insur-tech" services. Following the same concept of technological innovation, the other sectors of the economy began to designate similar terms, such as the health sector, in which technological startups are called Healthtechs; in education, smart platforms are known as Edutechs or Edtechs.

With a focus on a sustainable economy, innovative and alternative ideas arise for the most diverse situations of nature preservation. The startups in this area of activity are called Eco-techs. Biotechs are technological startups in the biological sciences: molecular, cellular, biodiversity, reproduction, and genetics. The term agro-tech refers to technology that is sustainable in solving agricultural problems. Reg-techs are startups that use information technology to improve regulatory processes, expanding to any regulated business with a particular appeal to the consumer goods industry and reducing millions in fines for companies. The operational domains of Tech companies are varied and continue to expand with the progression of digital transformation, integrating concepts like blockchain, artificial intelligence, digitalization, the hub economy, and the sharing economy. These startups' unifying characteristic is their technology foundation, irrespective of their specific industry sector. They are all categorized as Tech startups. Hence, we have broadly defined technology-based startups as X-Techs, where the "X" represents the business segment variable.

Studies on the factors that drive the life cycle of startups are scarce, as it is an area of knowledge under construction, but some authors serve as a source (Arruda et al., 2014; Marcon & Ribeiro, 2021). Most studies on the factors that increase the life cycle of startups deal with the importance of training founders and their teams (Zellmer-Bruhn et al., 2021; Roche et al., 2020) and the existence of an ecosystem to support and develop startups (Amaral, 2019; Gazel & Schwienbacher, 2021; Cukier & Kon, 2018). Given the inherent high risks, startups face a significant likelihood of failure due to the necessity for frequent strategic pivots and market repositioning to adapt to the dynamic environments in which they operate. Furthermore, early-stage ventures often grapple with limited financial, technological, and human resources while facing intense pressure to deliver results, particularly from investors (Kerényi et al., 2018; Kon, 2021). The scarcity of information connecting the various variables within the startup ecosystem, such as business segment, target audience, income model, and business stage, exacerbates uncertainties and anxieties for aspiring entrepreneurs.

Díaz-Santamaría and Bulchand-Gidumal (2021) studied the success of technology startups, and their finding suggests that four key factors significantly impact both measures of success: the startup's location, the dedication of its partners, the company's age, and the presence of non-promoting partners. Espinoza et al. (2019) point out the success of Chilean startups to the presence of universities and local patenting capacity. Alami et al. (2024) applied machine-learning methods to predict the failures of startups in Morocco and emphasized the critical role that capital and financial resources play in fostering business development. Although no single element can assure the triumph of a startup, founders who adeptly navigate and harmonize these factors are better positioned to establish an enduring and prosperous enterprise (Cukier & Kon, 2018; Honjo & Kato, 2019).

The general research question is: What are the key factors to the survival rate of Brazilian technologic-based startups? The initial hypothesis is that the X of the X-techs represents the business segment as a determinant factor to business survival.

The current study aimed to investigate the factors contributing to the business survival rate of X-Tech startups in Brazil. We examined the business segment, target audience, income model, and business stage variables using data mining techniques. The study's scope was limited to the Brazilian context due to the nature of the available data amenable to data mining analysis. By focusing on variables influencing startup survival, we aim to contribute to understanding these enterprises' dynamics in Brazil.

Startup Scenario Description in Brazil

The change in strategy and business paradigms has characterized startups' rapid and accelerated evolution (Kon, 2021). Combined with the world scenario, Brazilian technology-based startups (X-Techs) have been showing themselves as essential business models in contemporary economies, which can be observed both in academia and the market. In 2019, more than 12,000 startups were identified in the country, thus configuring an average annual growth in the number of units of 26.75% in the last nine years (Carrilo, 2020).

A study on the Brazilian startup ecosystem was carried out jointly by the Brazilian Startup Association—ABSTARTUP, Accenture Consultant, and FINEP (a public company linked to the Ministry of Science and Technology and Innovation). Results revealed that in 2017, about 73% of Brazilian startups were located in 10 communities; 63% had teams with less than five people; 46% had been working for less than two years; 41% were looking for traction; 44% were operating with the service model (SaaS) and 69% with an annual turnover of less than BRL 50,000 (ABSTARTUP 2021). Therefore, the concentration in large centers, the reduced size of the work team, and the short life span can be pointed out as characteristics of this type of company in Brazil.

According to ABSTARTUP (2022), using the Startup Base powered by the startups' inventors, which has become the largest statistical database on Brazilian startups, Brazil currently has 14,220 registered startups operating in all states and the Federal District. Although this number includes some companies already out of operation, results indicate the ideas that started and the evolution of this market in the country. This number of startups is segmented according to the segment of activity. Table 1 explains the segments that have the highest number of registered companies.

This profusion of startup technology promises multiple benefits, such as increasing efficiency and reducing costs for businesses or consumers as end users. Table 2 describes the X-Tech segments that predominate in the market segments.

Research Method

Data on digital transformation and startups are very complex and decentralized, mainly due to the absence of a history and a base unified and structured database. Thus, gathering different information to extract statistical data is necessary, which traditional methods could not process, extract, and analyze.

Previous studies suggest using data mining to extract information from a database (Gonzalez et al., 2016; Strang & Sun, 2020). The data mining process of uncovering patterns within statistical data (Han et al., 2012) provides insights that may remain obscured by traditional exploratory data analysis methods (Hand, 2000). Machine learning techniques, encompassing data mining, have emerged as a valuable instrument for identifying and investigating patterns and interdependencies among numerous variables.

Data mining applications offer classification models in some research areas, including health diagnosis and prognosis and identifying gaps in education data (Baker et al., 2011). Many forecasting methods use various data mining techniques. As computer processing performance has improved and various models have been proposed, research is being conducted to help entrepreneurs and investors find hidden patterns by applying data mining techniques (Kannan et al., 2010; Kim, 2021).

Table 1 Number of startups in Brazil by business segment

Business segment	Number of startups
Education	821
Others	709
Finance	567
Health and wellness	511
Internet	489
e-commerce	413
Agribusiness	348
Communication and media	334
Retail/wholesale	326
ICT and telecom	286

Source: ABSTARTUP (2022)

Table 2 X-Techs description by each segment

X-Techs	Description of startups by the segment of the market
ADTECH	Startups that develop technology for media convergence, big data analytics, and service distribution in decentralized programmatic media
AGRITECH	Startups that use technology focused on agricultural systems
AGROTECH	Startups that are focused on technological solutions for agriculture productivity
BIOTECH	Technology-based startups that work in the area of biological sciences
CLEANTECH	Startups that use technology to improve business performance, optimize processes, reduce waste and costs, pollute less, and reduce tailings production
ECOTECH	Startups technology-focused on a sustainable economy
EDTECH	Startups that use technology to scale education and promote accessibility to teaching and knowledge distribution
FINTECH	Startups that use technology to improve financial services
FOODTECH	Startups that offer technological solutions for the food sector, from delivery to the production process
GOVTECH	Technological startups that propose solutions for modernizing internal and public administration issues
HEALTHTECH	Startups that use technology to improve healthcare and wellness
HRTECH	Startups that facilitate and accelerate all stages of technology-based job hiring
INSURTECH	Startups focused on the insurance sector
LEGALTECH	Startups are dedicated to different functions to automate processes with artificial intelligence and accelerate the legal system's access and analysis of data
MARTECH	Startups that use technology to revolutionize digital marketing using bots, algorithms, big data, and data analytics
REGTECH	Startups that use information technology to improve regulatory processes
RETAILTECH	Startups that use technology to improve retail operations

Source: The authors

Data Pre-Processing

In general, collecting and pre-processing data analysis is essential and time consuming. For numeric data, relatively simple pre-processing is required, such as null value removal and categorical variable processing. However, there is somewhat complicated pre-processing for textual data, such as parsing, stopword elimination, and tagging. Most real-world data collected contains null values, and if any variable has multiple null values, the variable can be removed to allow parsing. Due to the lack of a consolidated base on startups, data from the online platform Startup base was used, called the official database of the Brazilian startup ecosystem of ABSTARTUP, and fed spontaneously by the startups' owners.

Initially, the X-techs were filtered by the activity segment (Table 1), totaling 12,907 registered in the different stages. Before using a large amount of data in forecast models, we applied dimensionality reduction methodologies to reduce the complexity time of the models, excluding some variables. At this stage, it was identified that the relevant segments in the country with the highest number of startups are Fintechs, Edutechs, Insurtechs, Healthtechs, Ecotechs, and Biotechs.

We excluded too many X-techs from our base, and this selection resulted in 2249 startups.

Upon deepening the investigation, it was found that within the area mentioned above of X-Techs, the company will focus on a segment: finance, education, marketing, well-being, health, and insurance. Most X-Techs are generally segmented into the market area (Fintech focused on finance, and Edutech focused on education). However, there is also a Fintech focused on education solutions, a Fintech focused on marketing, or an Insurtech focused on insurance finance. The three most relevant variables were selected in quantities of the selected X-Techs, resulting in the variation abbreviated to enable the processing (Table 3).

Three areas were selected from four areas in all X-techs except in Edutech. This occurred because the number of startups in advertising and the environment was the same, and in order to avoid undue selection, the authors decided to include both areas. After this categorization of the X-techs' areas of activity, the variable target audience income model and survival stage corresponding to each abbreviation were included in the spreadsheet (Table 4). In addition, descriptive stages such as start, traction, and business stage were summed up and presented in operation, as they are the open and available startups in the market (Table 5).

Figure 1 indicates the schematic of the information flow used when building the data mining array. We are looking for the causes of business survival (dependent output variable), considering four causable independent variables.

Table 3 Market area of the X-Techs and their abbreviation

X-Techs	Area of operation	X-Tech abbreviation
Fintech	Finance	FF
Fintech	Advertising	FA
Fintech	Education	FE
Insurtech	Insurance	IS
Insurtech	Finance	IF
Insurtech	Well-being	ISBE
Healthtech	Well-being	HSBE
Healthtech	Education	HE
Healthtech	Finance	HF
Edutech	Education	EE
Edutech	Finance	EF
Edutech	Advertising	EA
Edutech	Environment	EMA
Biotech	Biotechnology	BB
Biotech	Well-being	BSBE
Biotech	Agribusiness	BA

Source: Startup dataset. Conceptualized by the authors

Table 4 Independent variable values include target audience, income model, and business stage

Variable	Options	Definition
Target audience	B2B	Business to business
	B2B2C	Business to business to consumer
	B2C	Business to consumer
	P2P	Peer to peer
	B2S	Business to social
	B2G	Business to government
Business stage	FO	X – Tech out of operation
	In operation	X- Tech in operation
	Ideation	Idealization of X-Tech, business definition, and organization
Income model	SaaS	Software as a service is based on offering technological solutions through access
	MP	A marketplace is a business model that connects an offer with demand and monetizes that connection between the parties
	Other	Other business models
	E-C	A term refers to marketing products and services online, where transactions are carried out via electronic devices, such as computers, tablets, and smartphones. This type of commerce can occur through virtual stores, marketplaces, social network sales, and even email marketing
	Consumer	It develops products and services to serve the final consumer
	API	Application Programming Interface is a set of instructions that determines a communication pattern between two different software so that they can exchange information

Source: Gonçalves and Gonçalves (2021)

Table 5 Independent variables values: target audience, income model, and business stage in the business segment of the X-Tech FO, out of business; MP, target audience; E-C, e-commerce; API, application programming interfaces

X-TECH'S	Target audience						Business stage			IncomeModel					
	B2B	B2B2C	B2C	P2P	B2S	B2G	FO	In operation	Ideation	SaaS	MP	Other	E-C	Consumer	API
FF	144	99	63	9	2	1	41	298	28	120	67	70	13	29	8
FA	12	2	0	0	0	0	1	5	8	0	13	0	0	0	0
FE	4	3	2	0	0	0	4	4	0	3	2	0	2	2	0
IS	14	19	6	1	0	0	2	35	3	12	9	12	5	0	1
IF	3	5	4	2	0	0	1	11	2	6	4	0	1	1	1
ISBE	3	1	1	0	0	0	1	5	0	2	1	1	0	0	0
HSBE	99	87	73	2	0	5	46	246	50	111	58	58	13	16	3
HE	4	1	4	0	0	0	3	5	2	5	0	3	1	0	0
HF	4	1	0	0	0	1	0	6	0	3	0	1	1	1	0
EE	120	94	122	12	3	7	135	281	56	307	56	110	35	43	5
EF	4	5	10	2	0	1	3	17	3	4	3	5	1	8	1
EA	5	3	0	0	0	0	1	4	3	0	6	1	0	0	0
EMA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BB	8	8	4	0	0	0	3	15	4	1	2	11	3	2	0
BSBE	3	1	1	0	0	1	0	5	1	1	0	2	0	1	0
BA	4	0	0	0	0	0	1	2	1	1	1	0	0	0	0

Source: Startup dataset. Conceptualized by the authors

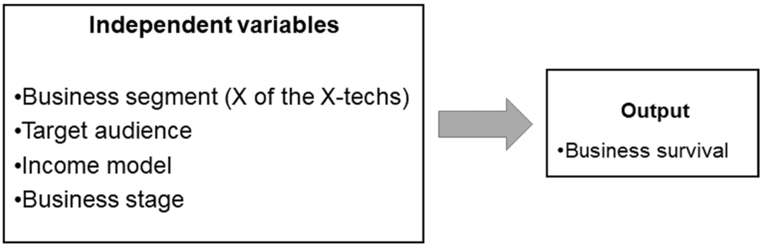

Fig. 1 Schematic of the information flow for building the data mining array

We used RapidMiner® (RapidMiner, 2017), a Java-based end-to-end analytical tool for data mining, text mining, predictive analytics, and business analytics (Almeida et al., 2016). This solution has been used in many areas and is the most popular standalone and open-source solution on the market (Poll, 2002) and the market leader in its field (Idoine et al., 2018). The optimal subset of variables is obtained by providing feedback on predictive performance until some condition is met using predictive models.

The output variable "business survival" was calculated considering the number of started X-Techs and the number that remains active (Table 6). To proceed with the data mining, we discretized the percentage found in a set of rules shown in the trees.

Table 6 Output values: The business survival of the studied techs is still operating. H, high; A, average; L, low

X-Tech's (business segment)	X-Tech started	Active	Business survival rate (%)	Discretization
FF	510	298	58.43%	A
FA	14	5	35.71%	A
FE	13	4	30.77%	L
II	51	35	68.63%	A
IF	18	11	61.11%	A
ISBE	6	5	83.33%	H
HSBE	435	246	56.55%	A
HE	12	5	41.67%	A
HF	10	6	60.00%	A
EE	768	281	36.59%	A
EF	23	17	73.91%	H
EA	8	4	50.00%	A
EMA	1	0	0.00%	L
BB	30	15	50.00%	A
BSBE	7	5	71.43%	H
BA	4	2	50.00%	A

Source: Startup dataset. Conceptualized by the authors

For discretizing the business survival into the levels high (H), average (A), and low (L), we used a rule described as if the "business survival rate" ≤ 30%, then "the survival rate" is low (L). If the "business survival rate" > 30% and ≤ 69%, then the "business survival rate" is average (A). If the "business survival rate" > 70%, then "the business survival rate" is high (H). The overall business survival rate referred to each X-Tech is shown in Table 6.

Data Mining and Analysis

The entire dataset was utilized to construct a random forest model using RapidMiner® Studio, a Java-based, open-source software (version 9.2, RapidMiner, Inc., Boston, MA, USA). We adopted the random forest algorithm for data analysis primarily due to its strengths in handling complex datasets, particularly when multiple variables and non-linear interactions are involved. Random forest is well suited for datasets with many variables (features). In the context of the present study, which involves analyzing the survival of startups based on multiple factors like income model, target audience, and business stage, random forest can efficiently handle this complexity as an appropriate choice (Lukita et al., 2023).

The analysis focused on predicting the "business survival" of X-Tech companies. The operators employed in the analysis included "retrieved data," "split data," and "random forest." The algorithm was trained using 80% of the dataset, with the remaining 20% reserved for model development. The model was built focusing on shuffling, pre-pruning, and the information gain ratio. This approach enhances the precision of classification by refining the attributes that distinguish between different samples in the training set (Lavrač et al., 1999). The accuracy was calculated using Eq. (1).

$$Accuracy(\%) = TP + TNTP + FP + FN + TN \times 100 \tag{1}$$

where TP = true positives, TN = true negatives, FP = false positive, and FN = false negatives.

In the context of a random forest model, accuracy provides a general indication of the model's effectiveness in making correct predictions. High accuracy suggests that the model performs well across the dataset, but it may not account for imbalances in the data or the distribution of classes.

Chohen kappa (κ) is a statistical coefficient of inter-rater reliability applied to evaluate two appraisers' agreement. κ is adjusted for the possibility of agreement occurring by chance. It compares the observed and expected accuracy if predictions were made randomly. In the present study, we accepted that the classification was appropriate when $\kappa \geq 0.60$. Information and data flow are shown in Fig. 2.

Results

On average, each segment has around 300 startups. These startups' most common target audience is B2B, and API is the most frequently used income model. Descriptive analysis revealed that 17.01% of the startups are related to education, 11.30%

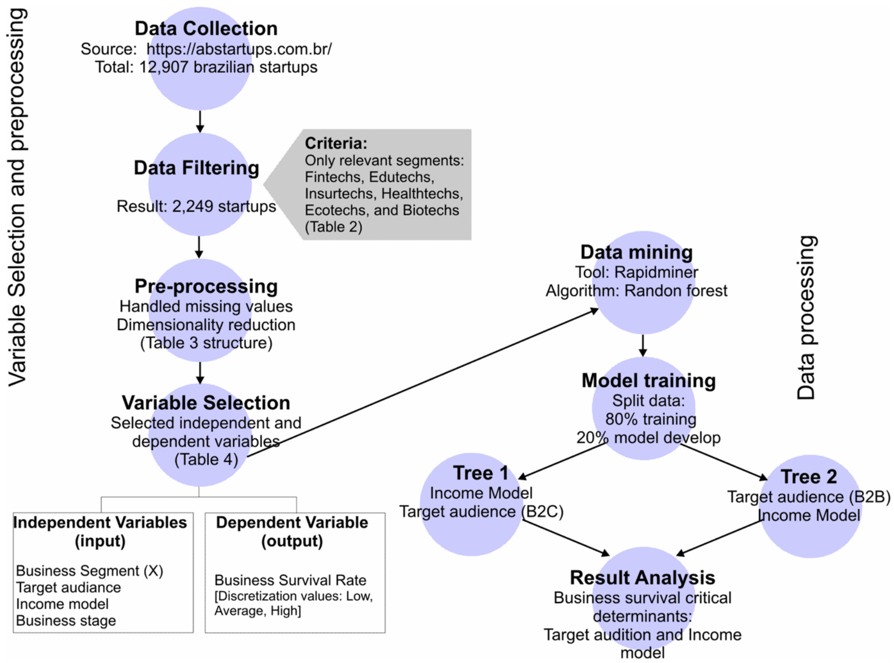

Fig. 2 Schematic representation of the adopted method

are in finance, 10.59% are in Internet infrastructure, 9.64% in well-being and health, and 8.68% in agribusiness. The remaining are in other fields (51.46%). In the education segment, 20.71% of the startups are out of business, 17.61% are in the welfare and health field, and 9.76% are in the finance area. The mean out-of-business rate across all startup sectors is approximately 18%.

Using the operator "random forest," we obtained two trees with an accuracy of 70% and $\kappa = 0.72$.

Tree 1

Figure 3 shows that the classification tree indicates that the target "business survival" was based on the "income model" such as the "target audience." This model refers to a digital platform where companies sell third-party products (for example, Amazon, Expedia, and OLX). Such a classification indicates an enhancement of transactions through the platform that benefits demand and supply. Rules were extracted from the tree-ensemble graph and presented in Fig. 3.

If the "target audience" is ≤ 1, then the "business survival" is "high" (representing 23% of the studied sample). This may indicate that if resources are invested in diversifying the target audience, such an action may lead to failure. If the "target audience" > 1, then we need to check the relationship between the company and consumers (B2C). If the "B2C" > 3, then "business survival" is "average" (46% of

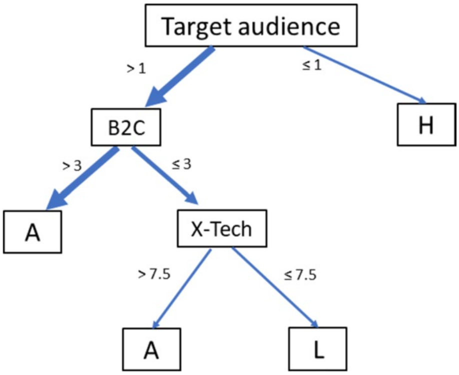

Fig. 3 The tree-ensemble graph based on the "target audience" attribute

the studied sample). If "B2C" ≤ 3, then we need to check the number of companies working in this "X-Tech" business area. If the total number of companies "X-Tech" is in the market ≤ 7.5, then the "business survival" is "low" (15% of the studied sample). If the total number of "X-Tech" ≤ 7.5, then the "business survival" is "average" (representing 15% of the studied sample). This indicates that X-Techs' survival rate is relatively low in Brazil.

Tree 2

Figure 4 shows that the classification tree indicates that the target "business survival" was based on the "B2B" consumer model. This model refers to a target audience that makes business directly to other businesses involving companies already well established in the market. The values involved are also higher than individual ones (for example, SalesForce, HubSpot, and Xero). The classification indicates a more significant step as the transactions when the principal customer is a business through the platform as beneficial to demand and supply. Rules were extracted from the tree-ensemble graph and shown in Fig. 4.

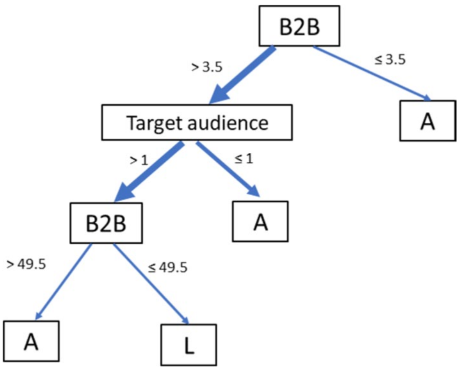

Fig. 4 The tree-ensemble graph based on the "B2B" root attribute

If the target audience "B2B" (business to business, consumer model) is ≤ 3.5, then the "business survival" is "average" (representing 15% of studied samples). If "B2B" < 3.5, one must check the "target audience." If the "target audience" ≤ 1, then the "business survival" is "average" (representing 38% of studied samples). If the "target audience" is > 1, then one must check on the B2B. If "B2B" ≤ 49.5, then the "business survival" is "low" (representing 31% of the studied samples). If "B2B" > 49.5, then the "target audience" is "average" (representing 15% of studied samples).

Discussion

The patterns shown in the trees tend to be more assertive as they consider the primary and secondary factors. In the analysis of the first tree-ensemble graph, we identified that the main factor of the survival of startups is the income model, with the target audience being the primary model for the longevity of startups. The choice of this income model already represents the survival of startups in 23% of the sample. After choosing the income model, the second variable to be considered is the target audience, with B2C being the primary focus for the survival of companies. Startups showing an average business survival rate represented 46% of the startups surveyed. This latest finding was unsurprising, considering that most startups with an income model target the B2C audience. If the survival rate of this B2C market analysis is less than or equal to 3, X-tech's operating segment must be considered. We can conclude with the analysis of the first tree that the primary variable to be considered for the longevity of a startup is not the market segment but the income model.

According to Cantamessa et al. (2018), many startups fail because they do not have a well-defined business development strategy. Such a move often leads to a lack of focus, misallocating resources, and, ultimately, failure to scale effectively. Early-stage startups, especially in the software industry, often fail because of inconsistencies between managerial strategies and their execution. This includes rushing products to market without adequate validation or neglecting the necessary learning processes (Giardino et al., 2014).

The second tree-ensemble graph identified that the primary startup business survival variable is the target audience, with B2B being the main focus for the survival of startups. The choice of this audience represents startup survival in 15% of the sample. After choosing the target audience, the income model is the second variable to consider. The leading income model is related to the survival of startups, presenting average business survival and representing 38% of the surveyed X-techs. In this tree, the results did not consider the startup's operating segment to reach the business survival; the operating segment was irrelevant. According to Goswami et al. (2023), capital scarcity and inadequate sales and marketing strategies toward the target audience are the most common factors for startup failure in India, similar to Brazilian startups.

The results using data mining analysis indicate that the greatest survival of startups is concentrated in the following:

1. **income model** ⇒ marketplace, and as a secondary variable the **target audience** ⇒ B2C;
2. **target audience** ⇒ B2B, and as a secondary variable, the **income model** ⇒ marketplace.

The marketplace income model offers entrepreneurs visibility, cost-effectiveness, security, and convenience. It serves as a virtual storefront where startups can showcase their products and services alongside offerings from various other companies. The marketplace provides customers with diverse products, like a digital shopping mall. This model is particularly advantageous due to the flexibility it affords companies, regardless of their size, to establish an online presence and the ease with which they can manage their virtual storefront. The negative factor is the competition for the attention of consumers, which is quite fierce. As a large online market, the B2C public target is the most common in the income model marketplace and B2B, as shown in Table 4.

The results of the present study partially agree with those of previous studies. Sevilla-Bernardo et al. (2022) found that the business model, the marketing approach, and the entrepreneurial team are the key factors for the success of startups. Silva Júnior et al. (2023) identified critical success factors influencing startup competitiveness, which included organizational, human, and environmental factors, such as internal characteristics, human capital, and the broader startup context.

The variables presented in this research, whether primary or secondary, testify that startups are ventures focused on technology and innovation, as previously proposed (Barbosa & Ramos, 2021; Donda, 2020). Furthermore, adopting the marketplace income model can enhance the survival rate of startups operating in volatile markets. Marketplaces provide increased exposure for products and services, showcasing them within a platform that inherently attracts customers independent of the specific businesses listed. Analogous to the established brand recognition of a physical shopping mall, the marketplace holds a solidified virtual brand presence online. Some examples of the consolidated marketplace in Brazil are Mercado Livre, Amazon, B2W, Shopee, Uber, and Air BNB. In addition to the ease of disclosure, another positive point of the marketplace is the possibility of rapid return, another essential feature for startups, agreeing with current literature on economic performance indicators (Westerman et al., 2012; Hess et al., 2016; Boneva, 2018; Heavin & Power, 2018).

In the present study, in addition to the income model marketplace variable, the target public variable was also crucial for the survival of startups, whether B2B (business to business) or B2C (business to consumer), as described by Gonçalves and Gonçalves (2021). Both B2B and B2C have large audiences with many consumers. They are the target audience of most marketplace platforms, enabling faster growth for startups, an essential item for consolidating startups in their segment (Kerényi et al., 2018; Kon, 2021).

The operating segment was found only in the third stage of analysis for the business survival, indicating that it is not a decisive perspective for the survival and advancement of startups to the following stages. From the 224 input variables used in this study, one can predict an improvement in the performance of the survival prediction model based on these perspectives by improving the future performance of startups with the development of prediction models. This structure can be considered a helpful tool to support decision-makers.

Relying heavily on marketplace platforms can offer significant benefits to startups, such as immediate access to a broad customer base, reduced overhead costs, and streamlined operations. However, this reliance also comes with several challenges and limitations that can impact a startup's survival and long-term success. Most of the startup's income might come from a single or few marketplace platforms. This concentration increases risk; if the platform changes its policies, increases fees, or if the startup is removed or demoted in the platform's search results, the impact on the business can be devastating. Regarding platform control, the marketplace controls vital aspects of the customer relationship, including data ownership, pricing, and branding. Startups have limited control over how their products are presented and marketed on these platforms. Startups may find it challenging to stand out among numerous competitors, many of whom may offer similar or identical products at lower prices.

Another point is that customers on marketplace platforms are typically looking for the best deal rather than brand loyalty, making it difficult for startups to build a loyal customer base. Marketplaces often restrict access to detailed customer data, limiting the startup's ability to gain insights into customer behavior and preferences, which hinders efforts to improve products and marketing strategies. While marketplace platforms offer significant opportunities for startups, especially in their early stages, over-reliance on these platforms can pose substantial risks and limitations. Startups must carefully balance the benefits of marketplace platforms with strategies to build their brand, diversify income streams, and maintain control over customer relationships and data.

The present analysis study documents two main findings that justify further research. An analysis of the survival of Brazilian startups was carried out using a data mining structure from the following perspectives: operating segment, income model, target audience, and business stage. We believe the results may contribute to the startup area's scholars and professionals and add to the emerging literature on startups.

In academia and the market, the doubts and uncertainties about startups are immense, and the findings initiate discussions to expand knowledge and reduce the startup survival risk. In the practical field, the main contribution of the results is to allow an analysis before starting a startup, focusing on the variables' target audience and income model instead of initially focusing on the X of X-tech, the operating segment. The prediction models for the specific domain structure on the startups' survival were built using four representative simple and fast regression algorithms. Using the attributes of performance segment, target audience, business stage, and income model, we reinforced the idea previously indicated by Steinbach et al. (2021) and Higuchi and Maehara (2021) on future consumer preferences.

When constructing the present study, we faced limitations in the existing information. We used a single database due to the scarcity of an official platform with statistical data on Brazilian startups. Future studies suggest applying the technique to other databases and countries to expand our knowledge about startup survival.

Virtual sales platforms, once a trend, especially during and after the COVID-19 pandemic, have become essential for any company and have many tools to optimize this e-commerce. The marketplace benefits small entrepreneurs who do not promote their products and services without spending a lot. Furthermore, among the most common difficulties in starting a startup are the lack of knowledge about the market in which it operates, turnaround time, and inference. Thus, the prediction models for the specific domain structure on the life cycle of startups in analyzing their survival were built using four representative simple and fast regression algorithms: segment, target public, stage, and income model.

Conclusions

The present research pinpointed variables contributing to a higher startup survival rate. The findings suggest that selecting the target audience and income model is the most determinant factor. Furthermore, introducing the term "X-tech" in the startup lexicon has provided a novel categorization, enriching the theoretical understanding of these ventures. As a contribution, the paper proposes the term X-tech, where X represents the business segment of technology-based startups. The main finding is the refusal of the initial hypothesis: the business segment did not appear as a significant factor, so it is not the X that leads to survival in the X-techs context but the target audience and the income model. These findings benefit constructing concepts about startups, which are many and diffuse.

To practitioners, startup entrepreneurs, and investors, this research's main contribution is the importance of considering the target audience and income model definition in the definition of business model strategy. Table 4 shows the choice options. To policymakers, the results reveal the importance of broad area startup support and incentive programs instead of specific business segment (for instance, education, insurance or finance) programs since the business segment is not a crucial survivor factor.

Whether primary or secondary, the variables indicate that inserting a startup into a marketplace platform with a B2B or B2C target audience will be more conducive to survival, as shown in the tree-ensemble graphs (Figs. 3 and 4). The company can benefit from website advertising to increase demand and do so at a low cost of marketing investment. The startup can gain visibility without worrying about the high costs of advertising, maintaining its e-commerce or programming professionals, and boosting sales. Another benefit is the logistics provided by a marketplace, which is more practical and easy to create means of payment, transport, and product registration, maximizing the company's profit. The marketplace is a beneficial tool to take advantage of people's easy access to social networks, using it to attract new customers, whether B2C or B2B.

The main limitations of this study arise from the data source. Due to the lack of an official platform with comprehensive statistical data on Brazilian startups, we relied on a single database from a startup association. Similar studies could be conducted in other countries or regions. Additionally, other data mining tools could have been explored.

Ultimately, the data mining method only indicated which variables are relevant to startup survival, and the results cannot be considered definitive. For future research, we suggest in-depth investigations through surveys or multiple case studies to understand why the target audience and income model are more influential than other variables.

Data Availability Data is available upon request to the first author.

Declarations

Ethical Approval The study used online open-access data and complied with all the academic good practices.

Competing Interest The authors declare no competing interests.

References

Abstartups, A. B. de S. (2021). StartupBase - A base de dados do ecossistema de Startups. Retrieved June 12, 2022, from https://abstartups.com.br/pesquisas/

Alami, R., Stachowicz-Stanusch, A., Agarwal, S., & Al Masaeid, T. (2024). Anticipating failure: A comprehensive analysis of entrepreneurship dynamic factors using machine learning predictive models. *International Journal of Central Banking, 20*(1), 327–348.

Almeida, P., Gruenwald, L., & Bernardino, J. (2016). Evaluating open-source data mining tools for business. In: *Proceedings of the 5th International Conference on Data Management Technologies and Applications* (pp. 87–94). https://doi.org/10.5220/0005939900870094

Amaral, M. (2019). Entrepreneurial ecosystems and the diffusion of startups by Alvarez, S., Carayannis, EG, Dagnino, GB, & Faraci, R (Eds.), (Edward Elgar Publishing, 2018).

Arruda, C., Nogueira, V., Cozzi, A., & Costa, V. (2014). *Causas da mortalidade de startups brasileiras* (p. 33). Núcleo de Inovação e Empreendedorismo, Fundação Dom Cabral Publishing.

Baker, R., Isotani, S., & Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira De Informática Na Educação, 19*(02), 03.

Barbosa, C., & Ramos, P. H. B. (2021). As Agtechs e o Ecossistema de Inovação do Espírito Santo. *Revista De Empreendedorismo e Gestão De Pequenas Empresas, 10*(1), 1.

Boneva, M. (2018). Challenges related to the digital transformation of business companies. In *Innovation Management, Entrepreneurship and Sustainability (IMES 2018)* (pp. 101–114). Vysoká škola ekonomická v Praze. Retrieved May 10, 2021 from https://www.ceeol.com/search/chapter-detail?id=690762

Cantamessa, M., Gatteschi, V., Perboli, G., & Rosano, M. (2018). Startups' Roads to Failure. *Sustainability*. https://doi.org/10.3390/SU10072346

Carrilo, A. F. (2020). Crescimento das startups: veja o que mudou nos últimos cinco anos. Retrieved April 12, 2022, from https://abstartups.com.br/crescimento-das-startups

Cukier, D., & Kon, F. (2018). A maturity model for software startup ecosystems. *Journal of Innovation and Entrepreneurship, 7*(1), 1–32.

Díaz-Santamaría, C., & Bulchand-Gidumal, J. (2021). Econometric estimation of the factors that influence startup success. *Sustainability, 13*(4), 2242. https://doi.org/10.3390/su13042242

Donda, M. M. D. S. (2020). Startups do agronegócio (AgTechs) no estado de São Paulo: perfil inovativo e práticas da gestão do conhecimento. Retrieved June 21, 2022, from https://repositorio.unesp.br/server/api/core/bitstreams/f95f593e-4fe5-4c6d-965a-0c5fe0bb89fa/content

Dremel, C., Wulf, J., Herterich, M. M., Waizmann, J. C., & Brenner, W. (2017). How AUDI AG established big data analytics in its digital transformation. *MIS Quarterly Executive, 16*(2), 81.

Espinoza, C., Mardones, C., Sáez, K., & Catalán, P. (2019). Entrepreneurship and regional dynamics: The case of Chile. *Entrepreneurship and Regional Development, 31*, 755–767. https://doi.org/10.1080/08985626.2019.1565421

Gazel, M., & Schwienbacher, A. (2021). Entrepreneurial fintech clusters. *Small Business Economics, 57*(2), 883–903.

Giardino, C., Wang, X., & Abrahamsson, P. (2014). Why early-stage software startups fail: a behavioral framework. In Software Business. Towards Continuous Value Delivery: 5th International Conference, ICSOB 2014, Paphos, Cyprus, June 16–18, 2014. Proceedings 5 (pp. 27–41). Springer International Publishing. https://doi.org/10.1007/978-3-319-08738-2_3.

Gonçalves, K. L. F., & Gonçalves, R. F. (2021). Nomenclatures, terminologies and classification of startups: A multivocal literature review. *Research, Society and Development, 10*(4), e53510414052–e53510414052.

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics, 17*(1), 33–42.

Goswami, N., Murti, A., & Dwivedi, R. (2023). Why do Indian startups fail? A narrative analysis of key business stakeholders. Indian Growth and Development Review. https://doi.org/10.1108/igdr-11-2022-0136

Han, S., Dang, Y., Ge, S., Zhang, D., & Xie, T. (2012). Performance debugging in the large via mining millions of stack traces. In: *2012 34th International Conference on Software Engineering* (ICSE) (pp. 145–155). IEEE. https://doi.org/10.1109/ICSE.2012.6227198

Hand, D. J. (2000). Data mining: New challenges for statisticians. *Social Science Computer Review, 18*(4), 442–449.

Hatada, F. (2021). Digital transformation of the insurance market in Brazil: An exploratory study on business models. MSc dissertation, Graduate Program Administração de Empresasda Universidade Presbiteriana Mackenzie. Retrieved June 20, 2021, from https://adelpha-api.mackenzie.br/server/api/core/bitstreams/749f64a9-be09-41d5-85ba-a39700604cde/content

Heavin, C., & Power, D. J. (2018). Challenges for digital transformation–Towards a conceptual decision support guide for managers. *Journal of Decision Systems, 27*(sup1), 38–45.

Hess, T., Matt, C., Benlian, A., & Wiesböck, F. (2016). Options for formulating a digital transformation strategy. *MIS Quarterly Executive, 15*(2), 123.

Higuchi, A., & Maehara, R. (2021). A factor-cluster analysis profile of consumers. *Journal of Business Research, 123*, 70–78.

Honjo, Y., & Kato, M. (2019). Do initial financial conditions determine the exit routes of startup firms? *Journal of Evolutionary Economics, 29*, 1119–1147.

Idoine, C., Krensky, P., Brethenoux, E., Hare, J., Sicular, S., & Vashisth, S. (2018). *Magic quadrant for data science and machine-learning platforms* (p. 13). Gartner, Inc Publishing.

Ismail, S., Malone, M. S., & Van Geest, Y. (2016). *Organizaciones exponenciales*. Bubok Publishing.

Kannan, K. S., Sekar, P. S., Sathik, M. M., & Arumugam, P. (2010, March). Financial stock market forecast using data mining techniques. In: *Proceedings of the International Multiconference of Engineers and Computer Scientists, 1*:4

Kerényi, Á., Molnár, J., & Müller, J. (2018). Bank and FinTechs. *Economy and Finance: English-Language Edition of Gazdasag Es Penzügy, 5*(1), 86–97.

Kim, M. (2021). A data mining framework for financial prediction. *Expert Systems with Applications, 173*, 114651.

Kim, M., Kang, S., Lee, J., Cho, H., Cho, S., & Park, J. S. (2017). Virtual metrology for copper-clad laminate manufacturing. *Computers & Industrial Engineering, 109*, 280–287.

Kon, A. (2021). Economia política das startups brasileiras: Nova ordem em um cenário de turbulências. *Brazilian Journal of Political Economy, 41*, 611–632.

Lavrač, N., Flach, P., & Zupan, B. (1999, June). Rule evaluation measures: A unifying view. In: *International Conference on Inductive Logic Programming* (pp. 174–185). Springer, Berlin, Heidelberg.

Lukita, C., Lutfiani, N., Panjaitan, A. R. S. et al. 2023. Harnessing the power of random forest in predicting startup partnership success. *Eighth International Conference on Informatics and Computing* (ICIC), Manado, Indonesia, 2023, pp. 1–6. https://doi.org/10.1109/ICIC60109.2023.10381988

Marcon, A., & Ribeiro, J. L. D. (2021). How do startups manage external resources in innovation ecosystems? A resource perspective of startups' lifecycle. *Technological Forecasting and Social Change, 171*, 120965.

Matt, C., Hess, T., & Benlian, A. (2015). Digital transformation strategies. *Business & Information Systems Engineering, 57*(5), 339–343.
Mercandetti, F., Larbig, C., Tuozzo, V., & Steiner, T. (2017). Innovation by collaboration between startups and SMEs in Switzerland. *Technology Innovation Management Review, 7*(12), 23.
Poll, K. (2002). What main methodology are you using for data mining. Retrieved June 12, 2020, from www.kdnuggets.com/polls/methodology
RapidMiner. (2017). Data science platform. In *RapidMiner Studio (Version 9.0)*. RapidMiner, Inc. Retrieved May 10, 2020, from https://rapidminer.com
Roche, M. P., Conti, A., & Rothaermel, F. T. (2020). Different founders, different venture outcomes: A comparative analysis of academic and non-academic startups. *Research Policy, 49*(10), 104062.
Schwertner, K. (2017). Digital transformation of business. *Trakia Journal of Sciences, 15*(1), 388–393.
Sevilla-Bernardo, J., Sanchez-Robles, B., & Herrador-Alcaide, T. (2022). Success factors of startups in research literature within the entrepreneurial ecosystem. *Administrative Sciences, 12*(3), 102. https://doi.org/10.3390/admsci12030102
Silva Júnior, C. R., Siluk, J. C. M., Neuenfeldt-Júnior, A. L., Francescatto, M. B., & Michelin, C. D. F. (2023). Mapping risks faced by startup investors: An approach based on the apriori algorithm. *Risks, 11*(10), 177. https://doi.org/10.3390/risks11100177
Steinbach, J., da Fonseca Burgardt, V. D. C., de Castro-Cislaghi, F. P., Machado-Lunkes, A., Marchi, J. F., do Prado, N. V., ... & Mitterer-Daltoé, M. L. (2021). Understanding consumer, consumption, and regional products: A case study on traditional colonial-type cheese from Brazil. *International Journal of Gastronomy and Food Science, 26*, 100418.
Strang, K. D., & Sun, Z. (2020). Hidden big data analytics issues in the healthcare industry. *Health Informatics Journal, 26*(2), 981–998.
Von Leipzig, T., Gamp, M., Manz, D., Schöttle, K., Ohlhausen, P., Oosthuizen, G., ... & von Leipzig, K. (2017). Initialising customer-orientated digital transformation in enterprises. *Procedia Manufacturing, 8*, 517–524.
Westerman, G., Tannou, M., Bonnet, D., Ferraris, P., & McAfee, A. (2012). The digital advantage: How digital leaders outperform their peers in every industry. *Mitsloan Management and Capgemini Consulting, MA, 2*, 2–23.
Zellmer-Bruhn, M. E., Forbes, D. P., Sapienza, H. J., & Borchert, P. S. (2021). Lab, Gig or Enterprise? How scientist-inventors form nascent startup teams. *Journal of Business Venturing, 36*(1), 106074.

Authors and Affiliations

Kelly Lima Fonseca Gonçalves[1] · Berislav Andrlić[2] · Rodrigo Franco Gonçalves[1] · Irenilza de Alencar Nääs[1]

✉ Rodrigo Franco Gonçalves
rofranco212@gmail.com

Kelly Lima Fonseca Gonçalves
kellylfg7@gmail.com

Berislav Andrlić
bandrlic@ftrr.hr

Irenilza de Alencar Nääs
irenilza.naas@docente.unip.br

[1] Universidade Paulista, São Paulo, Brazil

[2] Polytechnic in Pozega, Pozega, Croatia