International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(9s) 2026 pp. 269-286
https://doi.org/10.70917/ijcisim-2026-3440
Article

Sustaining Momentum: A Data-Driven Analysis of Funding Gaps and Lifecycle Longevity in Startups

Lakshmi S[1], Jayalakshmi N.A[2] , Sindhu.T[3], Karuna M[4],Maithreyi Mahesh[5]
[1,2,3,4,5]Mount Carmel Deemed to be University, Email: lakshmi.raagu@mccblr.edu.in

Abstract: Startups are undeniably the drivers of tech innovation, job creation and economic growth, but how do they connect to capital? So, although it is undeniable that startups are the drivers of tech innovation, job creation and economic growth, how do they connect to capital? In reality, less than 90% of those who embark on growth reach long-term sustainability. Historically, The challenge has been to analyze why startups fail or succeed. The majority of studies are based on small area information or short timeframes or fixed linear models. Unfortunately, these Low-tech solutions do not capture the messy, non-linear dynamics that really creates. modern entrepreneurial ecosystems. We address these gaps by using a statistical, machine learning based approach which systematically explores the structural, financial and temporal determinants of startup performance. Combining a comprehensive dataset of 54,294 startups from 1902 to 2015 in 115 countries, we propose a cluster of startup outcomes as a multiclass property consisting of operating, acquired and closed. We validate different classification models: Logistic Regression, XGBoost, Multilayer Perceptron (MLP) and CatBoost through train and test split of 80–20 and performance on accuracy and Area Under the Receiver Operating Characteristic curve (ROC-AUC). Overall, the empirical results suggest the age at last funding is the leading driver for startup success — the drivers are, then as at this stage, sustainable investor confidence, operational maturity — they are essential for any new venture to have. Market sector, which became the second most important outcome predictor as well; sectors that have a technology-driven, highly scalable business model (software, fintech, e-commerce and biotechnology) are much more successful than capital-intensive and low-scalability industries. Additionally, the amount of venture capital invested is closely connected to overall survival and acquisitions and the number of rounds, and the gap between rounds of venture capital funding. Methodologically, CatBoost classifier is found to provide the strongest overall performance as it excels in handling complex categorical variables over high cardinality, extreme class imbalances and non-linear relationships without any target leakage. This information is a pragmatic guide for everyone involved. It's a matter of solving their fundraising problem for the founders, and their investing problem for the investors. better way to hedge bets. And for policymakers, it offers the very toolkit they require to create a startup ecosystem that can weather the storm
Keywords—Startup success, Machine learning, CatBoost, Venture capital, Entrepreneurial ecosystems, Predictive analytics, Startup failure.

1. INTRODUCTION

Startups globally embed themselves as mechanisms of radical innovation, market disruption, and structural economic transformation. These high growth organizations have a business model that embraces agility and innovation, their scalability targets are more intense than their early competitors operate in extremely uncertain and resource-scarce conditions. Startups are structurally fragile as they also have strategic value. Empirical evidence suggests that around 90% of new startups fail in their first few years of existence, a metric that has been consistently stagnant, despite record volumes of global venture capital injection, professionalized incubator services, and policy nudges to support new ventures. That failure rate endures demonstrates a profound inadequacy in knowledge concerning what actually makes a startup successful and what makes a startup fail. The definition of what constitutes a startup's 'success' has changed. Traditional economic literature often used short-term, speculator data points such as paper valuations, fast growth of top line revenues, or early time to entry on early-stage venture financing. Modern scholarship, however, is more often discarding these cyclical, volatile metrics in favor of broader, structural outcome-based indicators. Contemporary frameworks indicate the success of an organization through the continuing operational viability, continued survival of a structural structure, and the issuance of a company in order to exit its liquidity, whether through acquisition or public listing. However, success is not a set achievement, as far as this is concerned. A phenomenon that evolves over time because of the decisions made through the key strategies adopted, such as how capital is allocated. The elevation and deployment of raised and the geographic environment along with the company's stage of raised, were chosen. maturity. Earlier research in entrepreneurship largely relied on descriptive approaches or traditional linear models, including Ordinary Least Squares regression, survival analysis, and basic logistic regression frameworks. Although these approaches provide basic interpretability, they introduce restrictive parametric assumptions about how independent variables are, how distributions should be, and whether there are log-proportional hazards. Conversely, real-world empirical startup datasets are quite complex. They have high dimensionality, considerable multicollinearity, large number of missing entries, categorical dominance, and the extreme class imbalance that arises from high base failure rates. Consequently, conventional frameworks have limited predictive ability and an increased risk of applying biased statistical procedures. We now have an alternative in new predictive analytics and ensemble machine learning capabilities. Sophisticated gradient-boosting architectures have the ability to map nonlinear interactions mathematically; allow a heterogeneous feature space on the task or handle non-normal outcome classes; and handle the problems of class imbalance without the need to reduce data to their normal state using artificial data normalization. Empirical evidence demonstrates that in complex organizational environments, machine learning architectures always outperform classical linear baselines, as feature importance metrics are discovered that are not only theoretical but are based on practice in the business. By modeling a long-horizon large-scale historical cohort that includes 54,294 firms from 115 countries established over the years 1902 to 2015, this study sheds light on how startup success is decoupled from near-term market cycles. Going well beyond limited geographic boundaries or sector concentrations, this study applies ensemble machine learning, notably with the focus on the CatBoost framework, to develop a rigorous empirical ranking of structural success drivers.

2. CONCEPTUAL FOUNDATION

Inspired by the concepts from entrepreneurship, venture capital finance, organizational risk, and data driven analytics, the study conceptualizes startup success as a complex interaction of structural, financial, temporal, and environmental factors between interrelated and non-interrelated variables. Rather than treating the performance of a startup as a function of independent variables, this model regards success as an emergent consequence of cumulative signals created by the funding mechanics, market selection, geographic circumstances and maturity of the firm over time. There is an unpredictable operating environment for entrepreneurship in its true entrepreneurial framework in which startup's success relies on the discovery of attractive market opportunities, the reshaping of business models or resource acquisition through the challenging peer networks. Previous studies have emphasized that success is multi-faceted and also success should be judged on the performance-metrics defined outcomes such as survival, acquisition, and ongoing operations rather than short term valuation and growth metrics - which are mostly based on market cycles determined by speculative cycles (Sevilla-Bernardo et al., 2022; Kim et al., 2023). This study adopts a practical outcome based perspective and defines startup success as a measurable property, measured through persistence, scalability and strategic alignment. The theory is similarly driven by venture capital theory that illustrates the importance of framing and sequencing of financing. Staged investment is understood as a remedy for agency problems as well as uncertainty as it allows investors to assign capital using measurable performance intervals (Gompers, 2022). Other governance features embedded within venture capital architecture, e.g., monitoring, board oversight and control rights also shape strategic discipline and organizational learning, and can even enable a startup to succeed and even fail in the long run (Sahlman, 2022). Syndication among experienced investors magnifies these effects through the integration of expertise and reduction of idiosyncratic risk, particularly at the onset (Lerner, 2022). Therefore, funding rounds as well as sustained persistence of funding investment and venture capital engagement are considered core explanatory variables in the present study. Temporal dynamics are a key component of the conceptual base as well. The interval of time between two consecutive rounds of funding is a predictor of firm performance internally, as well as an indicator of confidence from external investors. Long funding gaps frequently are also a signal of operational and market instability, less growth or adverse market conditions associated with longer periods, resulting in higher probability of failure (Singh and Mungila Hillemane, 2023). On the other hand, both shorter and consistent funding intervals indicate that they are accompanied by momentum and validation that heighten the potential for survival. In this regard, age of the firm, including age at last funding, has accordingly been considered an accumulative measure of resilience and maturity within an organization. Risk theory brings even more focus on startup outcomes as it proves that entrepreneurial failure is not merely a single event of destruction in which a single risk factor renders the entrepreneur riskless. Financial, technological, operational and market risks interact jointly to produce viable companies, but also investors' perceptions of these risk factors influence access to finance, strategic options and so on (Fiet, 2022; Skrynkovskyy and Tyrkalo, 2021). At the macro level, macroeconomical shocks and business cycle oscillation have the long-run impact on the longevity of the business and investment behavior, thus we propose that models that consider structural as well as short-run factors influencing success are necessary (Gourio, 2012). Methodologically, this study focuses on the research of startup success from its perspective because traditional linear statistical models are not sufficient because their data may display problem such as non linearity, multicollinearity, high dimensional categorical variables, missing data, and high-class imbalance in entrepreneurial dataset (Chan et al., 2022). Machine learning methods, and especially ensemble and gradient boosting methods, overcome these limitations by understanding complex relationships, addressing imbalanced results and managing heterogeneous space among features (Zhang et al., 2022; Asselman et al., 2023). In this case, machine learning is more than that: Simplest prediction and is a contributor to theory validation. By analysing feature importance, becomes possible to identify and rank the most influential factors – such as funding, market demand, and legal and regulatory frameworks, etc. Identify and describe the type, location and maturity of the firm as well as linking them to other theories on entrepreneurship. finance (Bangdiwala et al., 2022; Kim et al., 2023). This is a way of thinking about starting-up success as if it were a tracing the cause of the results, which can be used to guide data-informed actions. decisions to the founders, investors and policymakers.

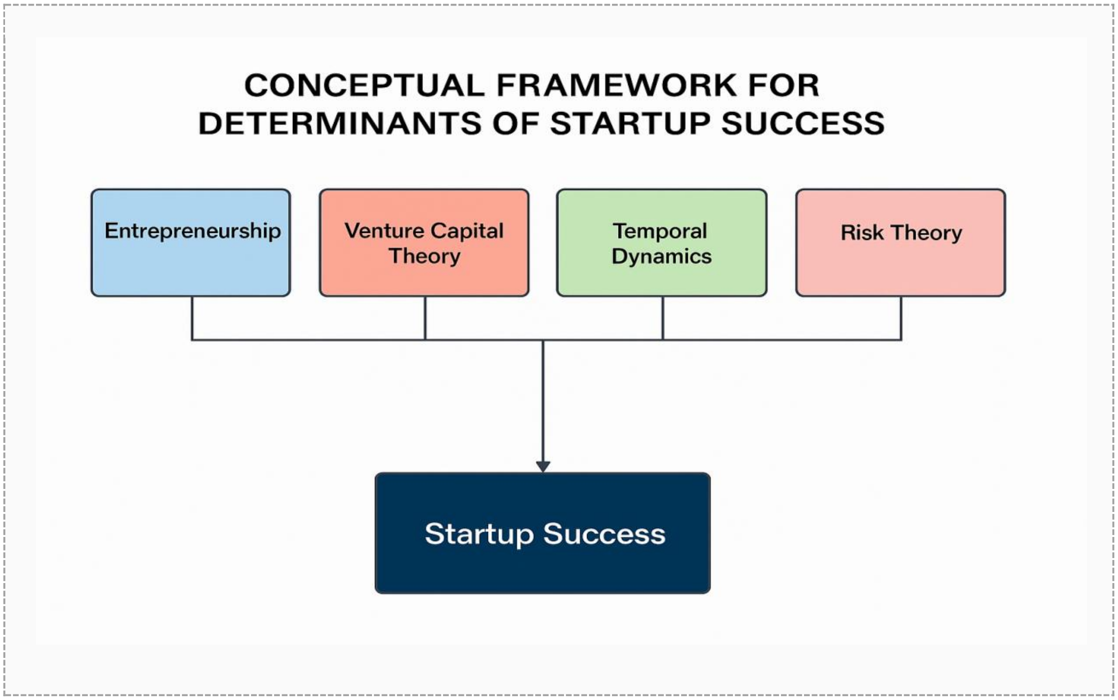

Fig. 1

2.1. Entrepreneurship and Market Selection

Startups' working conditions are extremely tough and resources have to be mobilized. They grew rapidly and had to constantly change to cope with life. This paper has a clear and observable results (i.e. whether firm is still in business or acquired). By focusing on these end states, the analysis avoids being influenced by short-term market hype or inflated valuations. Rather, it's about ongoing success, with robust strategic alignment and organisational framework that may be flexible for changing over time.

2.2. Venture Capital and Staged Financing Theory

Venture capital is the field that's most frequently associated with staged financing, which is employed to limit liability and match the interests of In particular, investors and founders in uncertain environments. In addition to investing all of his money in Capital will be unlocked in phases, each corresponding to a set of milestones. Along with Institutional investors can be involved in governance, often through the board, and enjoy the benefits of funding. For better decision making, participation, monitoring and control rights are being used. and accountability. Syndicated investments do much the same, too, by putting start-ups in touch with. Join wider networks, share risks and leverage on a variety of expertise. Combined, the structure The amounts and when funding rounds occur can give some indication of a company's future prospects.

2.3. Temporal Dynamics and Signaling Theory

A startup's timeline can be helpful when determining the overall performance of the startup. In particular, the The number of years between investments can indicate a firm's development. Regular and timely Typically strong operations and continued investor confidence are represented in the fundraising. In contrast, Extended periods between rounds can be a sign of internal struggles, missed targets or falling standards. Self-confidence, which can lead to a greater risk of limited resources. Consequently, the age of the company as well as the number of its employees are important factors. The number of times the company has funded a round and time since the last funding round are useful indicators of resilience and stage of the company. development.

2.4. Organizational Risk and Ecosystem Theory

As far as risk is concerned, failure in startups is seldom attributable to only one thing. It typically results From overlapping and overlapping risks. There are a variety of issues that startups have to face— Technological, operational, financial and market factors—are also addressed, and in tandem. Unpredictable events or economic situations. Traditional linear models can fail to adequately capture the structure in the data. These interactions can be captured, especially if there are interactions between variables or variables are dominant. In In contrast, machine learning techniques, particularly the gradient boosted models, are more appropriate for these types of problems. Recognizing complex patterns in different data types. Feature importance analysis is another tool that can be used to help. relate these patterns to practical, real-world insights.

3. REVIEW OF LITERATURE

Investigation of the success of startups has spanned entrepreneurship, venture capital, innovation management, and data analytics, with researchers attempting to answer the question of why few startups make it to the exit and survive (or get off the ground) whereas many start-ups can't. Primitive research focused attention on descriptive studies of founder characteristics, firm capabilities, and macroeconomic factors, and they focused much on characteristics of founders but often resulted in few empirical generalizations. Recent research increasingly considers ecosystem-level factors as well as firm specific factors as explanatory of entrepreneurial outcomes (Sevilla-Bernardo et al., 2022). Startup success is a well-known construct, but is not a singular performance measure. Many empirical studies show that survival, acquisition, as well as continued operational sustainability constitute a more consistent measure of success compared to valuation and short-term growth indicators that are usually biased by capital market cycles (Kim et al., 2023). At its core, a venture's success depends on how well its market choices align with its financing structure, governance practices, and stage of organizational maturity over time (Gompers, 2022). Research in venture capital consistently shows that both the structure and timing of funding play a major role in shaping a startup's growth path. Staged financing, for instance, can be available that permits investors and founders to Collaborate and act as a team, but have some control over access to funds, funds only released once important milestones are reached are achieved (Gompers, 2022). Governance mechanisms such as active investor monitoring Also, and board oversight has significant responsibilities for strategic decision making and oversight. As the firm expands, accountability is also increasing (Sahlman, 2022). In addition, well-informed VC's contribute not only capital, but also valuable expertise and networks, which can be particularly valuable. Singh and Mungila Hillemane (2023) demonstrate the cyclical nature of startup financing needs, suggesting that startup finance in its initial stages significantly relies on the human and research capital invested, whereas the latter phases largely depend on institutional investors and private equity for scaling up growth. Entrepreneurial risk has also been identified as one of the primary determinants of startup performance. Investors differ systematically in how they see and handle markets and agency risks, which influences funding decisions and how long a firm survives (Fiet, 2022). Data is then further categorized to include startup risk across financial, technological, operational and market based categories based on the finding that failure most often comes from the interplay of several different dimensions of risk and not from isolated shocks (Skrynkovskyy and Tyrkalo, 2021). This phenomenon is more pronounced at macro-levels, where rare economic disruptions have lasting impacts on firm survival and investment behavior across the business cycles (Gourio, 2012). Methodologically, early startup research, with its emphasis on structured research methods such as regression and survival analysis models, is dominated by traditional regression and survival analysis models, which remain weak for large entrepreneurial datasets. The startup data has non-linearity, has high dimensional categorical variables, multicollinearity, and missing values with extreme imbalance in classes due to failure frequency. These properties limit the applicability and prediction capabilities of traditional statistical models (Chan et al., 2022). In order to meet these challenges, machine learning techniques have been increasingly applied in recent studies. Ensemble and gradient boosting algorithms show the best performance in class with complex feature interaction (Zhang et al., 2022) and imbalanced outcomes. Asselman et al. (2023) demonstrated that boosting-based models are proven to outperform traditional approaches on accuracy and robustness in all of literature domains. Machine learning models for startup-specific settings have outperformed traditional approaches in predicting acquisition and survival outcomes (Bangdiwala et al., 2022). A hybrid model that integrates firm-level internal variables with external environmental variables provides the best predictive performance among these methods. Kim et al. (2023) indicate that the use of funding characteristics, market, and organizational maturity have positive effects on prediction of startup success. As a result, it indicates that machine learning has improved both the predictive power and explanatory capability, and feature importance analysis can act as a mediating variable between the theoretical and practical behavior. However, such studies suffer from time constraints, sector-based as well as geographic specialization. Historical longitudinal datasets covering entrepreneurial dynamisms across a wide technological and economic spectrum have received little attention. The present study bridges the gap in our line of research by using the machine learning approach to a large-scale startup dataset over a long-horizon, large dataset to detect structurally persistent determinants of success.

4. RESEARCH DESIGN AND METHODOLOGY

4.1. Research Design Overview

The data in this study is secondary data and is longitudinal data that was originally compiled by the data provider, Crunchbase, and obtained from the Kaggle data repository. Crunchbase is a complete business intelligence platform that monitors the enterprise life cycle from start-up to exit, real-time private funding data, and corporate governance activity around the world. The historical cohort of startups comprises 54,294 distinct startup entities formed across a multi-epoch period of time from 1902 to 2015 in 115 countries.

The structural data contains 39 attributes that represent qualitative properties of the firms, including official legal naming, primary market sector, specific operational indicators (operating, acquired or closed), and extremely fine-grained geospatial data such as ISO countries, state identifiers, regional borders and local city hubs. Moreover, the information features detailed quantitative columns, including cumulative capital amounts; chronological columns (date of incorporation, initial funding date, last funding date); and sequential tracking of institutional investment arrays from Series A through Series H. This comprehensive analysis offers a more holistic perspective on the nature of survival, over several economic cycles and technology waves.

4.2. Variable Operationalization

The dependent variable Startup Success is a categorical variable with three possible values: Operating, acquiring and closing. Firms that are still in business, or acquired, are treated. as successful, as they either still operate or have undergone a “liquidity event”, Closed firms are deemed to be failures. This definition minimises the impact on the changing The method of valuation and similar characteristics are consistent with the common method used in entrepreneurship research. The model seeks to integrate independent variables across four dimensions, ranging from funding metrics to its impact on the L2 community.It attempts to integrate independent variables in four categories: funding metrics, to impact on the L2 community. temporal, sector and geographical factors. These make up 39% of the total. In a structured feature, as described below.

TABLE I
Attribute Dictionary and Variable Operationalization of the Crunchbase Dataset

Sl. No	Name	Explanation
1	Permalink	Permanent URL pointing to the company's profile page
2	Name	Official legal/trading name of the startup
3	Homepage_url	Link to the company's primary website
4	Category_list	List of semi-colon separated relevant categories

5	Market	Core field of macroeconomic competition
6	Funding_total_usd	Total cumulative funding received in USD
7	Status	Target Variable: Current operational status (Operating, Acquired, Closed)
8	Country_code	Three-letter ISO code of home country
9	State_code	Two-letter code of home state (where applicable)
10	Region	Geopolitical code of home region
11	City	Code/Name of home city
12	Funding_rounds	Total integer count of funding rounds completed
13	Founded_at	Exact calendar date of formal founding
14	Founded_month	Calendar month of formal founding
15	Founded_quarter	Fiscal quarter of formal founding
16	Founded_year	Calendar year of formal founding
17	First_funding_at	Exact calendar date of first funding round received
18	Last_funding_at	Exact calendar date of last funding round received
19	Seed	Total capital raised in Seed stage rounds
20	Venture	Total capital raised from institutional venture capitalists
21	Equity_crowdfunding	Total capital raised via equity crowdfunding channels
22	Undisclosed	Capital raised from undisclosed financing sources

23	Convertible_note	Capital raised via short-term convertible debt instruments
24	Debt_financing	Total capital raised from formal loans and credit lines
25	Angel	Total capital raised from individual angel investors
26	Grant	Total non-dilutive capital received via grants
27	Private_equity	Capital received from late-stage private equity stocks
28	Post_ipo_equity	Public equity capital received after an initial listing
29	Post_ipo_debt	Debt financing capital received after an initial listing
30	Secondary_market	Funds raised via secondary equity market transactions
31	Product_crowdfunding	Capital raised from the public via product pre-sales
32-39	Round_A to H	Explicit capital volumes raised in specific Series A through H rounds

4.3. Data Preprocessing Pipeline

To prepare the dataset for analysis, a series of preprocessing steps were carried out:

- Missing numerical values were replaced with median values to limit the influence of outliers, while missing categorical entries were filled using the most frequent category.
- Date variables (e.g., founded_at and last_funding_at) were converted into more easily interpreted numeric variables (e.g., firm age, age at last funding, time between funding rounds).
- An attempt was made to scale the numerical features so that they were on a similar scale and not having a disproportionately large effect on the model.
- The dataset was then divided into training (80%) and testing (20%) subsets, enabling evaluation on unseen data and helping to reduce overfitting.

4.4. Model Development and Evaluation Architecture

Four classification algorithms were analyzed and selected for their respective reasons. A baseline was used to compare with Logistic Regression. To account for non-linear relationships, an ensemble tree approach was used with XGBoost, and a Multilayer Perceptron (MLP) was used to investigate more complex patterns in the data. The choice of the main model was made in favor of CatBoost, which is especially suitable for working with categorical data that has many different values, and is not so demanding on pre-processing. Two common metrics were used to evaluate model performance: classification accuracy and ROC-AUC.

5. EMPIRICAL RESULTS AND DATA ANALYSIS

5.1. Exploratory Data Analysis and Descriptive Statistics

When checking the target status variable of the dataset, the class imbalance is noticeable. If we take as our initial descriptive statistics, it is empirically proven that 86.9% of the startups in this historical database are classified as actively operating. In contrast, 7.7% successfully achieved liquidity exits via acquisition, and 5.4% are reported to be formally closed.

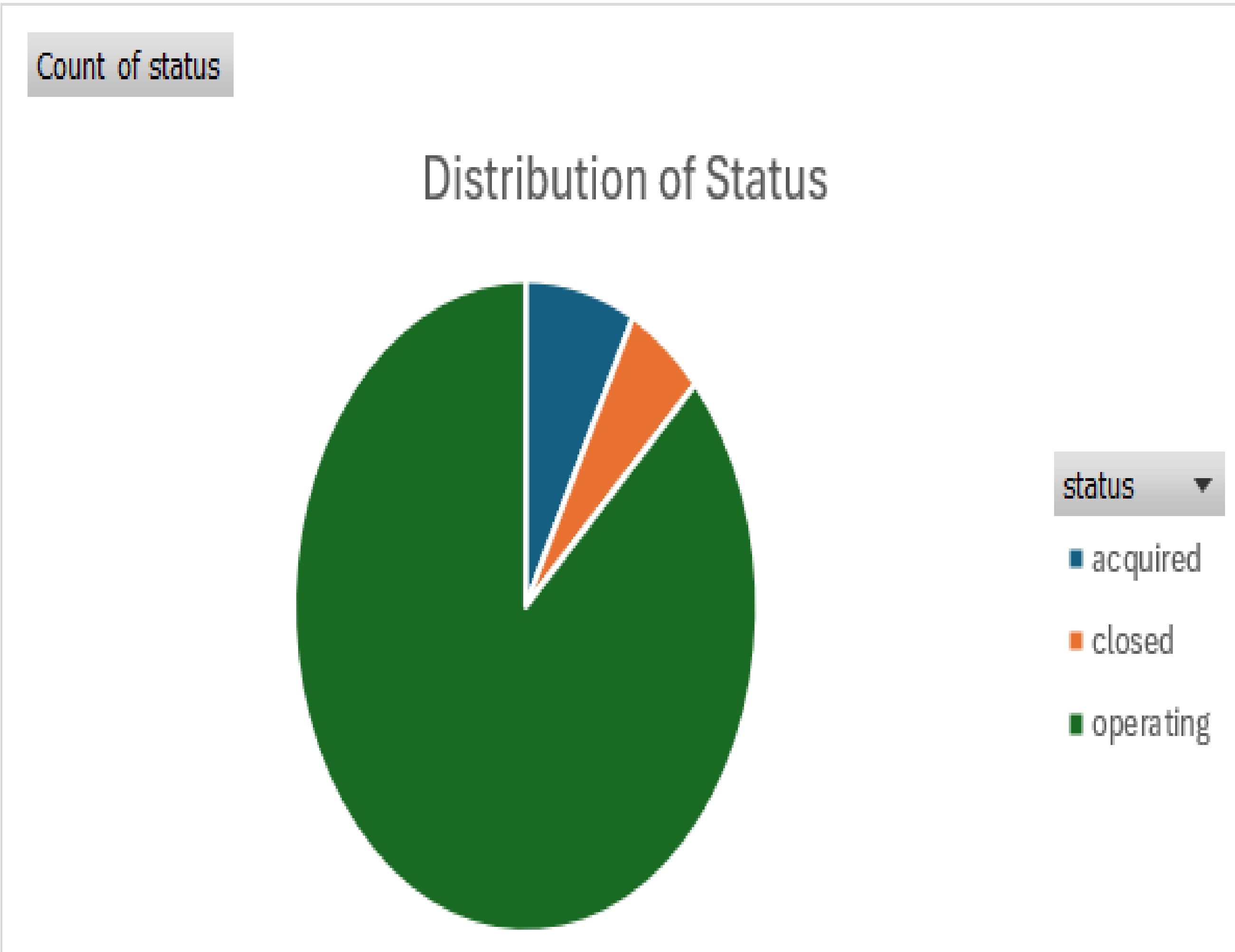

Fig. 2: *Empirical distribution of the target variable 'Status' displaying class imbalances across operating (86.9%), acquired (7.7%), and closed (5.4%) firms.*

From a market viewpoint, this high concentration of working companies implies that numerous early-stage and mid-stage companies postpone formal closure or bankruptcy. Founders can maintain active legal registrations in order to hold on to hard-to-recover corporate licenses, protect brand equity, or retain active contracts in an ideal scenario for future revival. Another good point is the option of maintaining a registered entity (rather than moving into formal liquidation) may cost less, and the prevention of a public closure notice can also protect an entrepreneur's professional reputation with future investors and business partners.

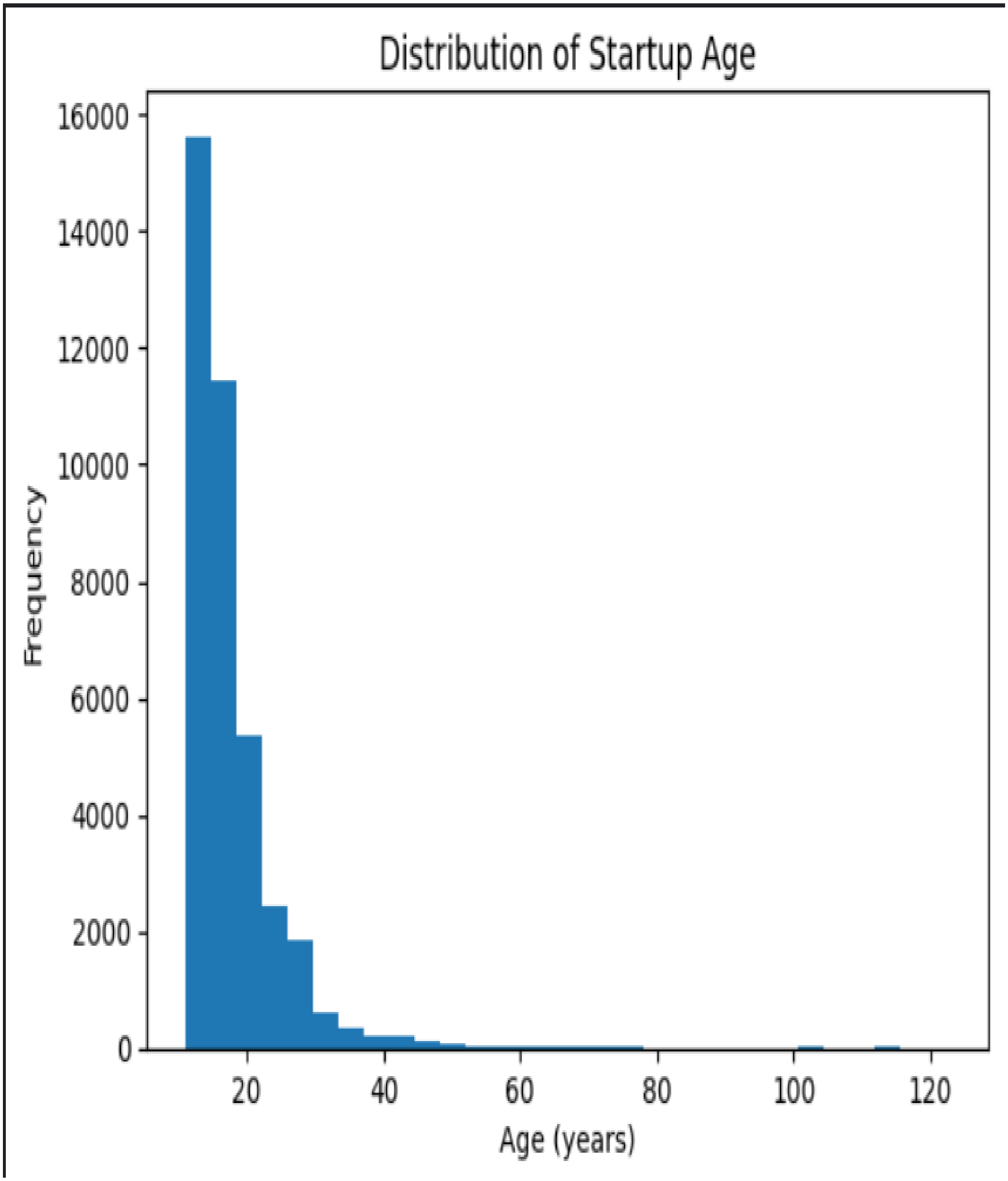

Fig. 3: *Distribution of Startup Lifespans Highlighting Early Failure Risks*

The figure below is a histogram of the length of startups. The majority of firms, in the early years between 0 and 20 years, are grouped together, giving a distinct right-skewed distribution. Following this period the number of companies rapidly declines and a small percentage outlast 100 years, creating a long tail. This trend indicates that the initial start of a start-up is particularly difficult as many companies don't make it through the first few hurdles of operating and the market. By contrast, the handful of companies that endure for a longer time demonstrate good resilience and adaptability to change.

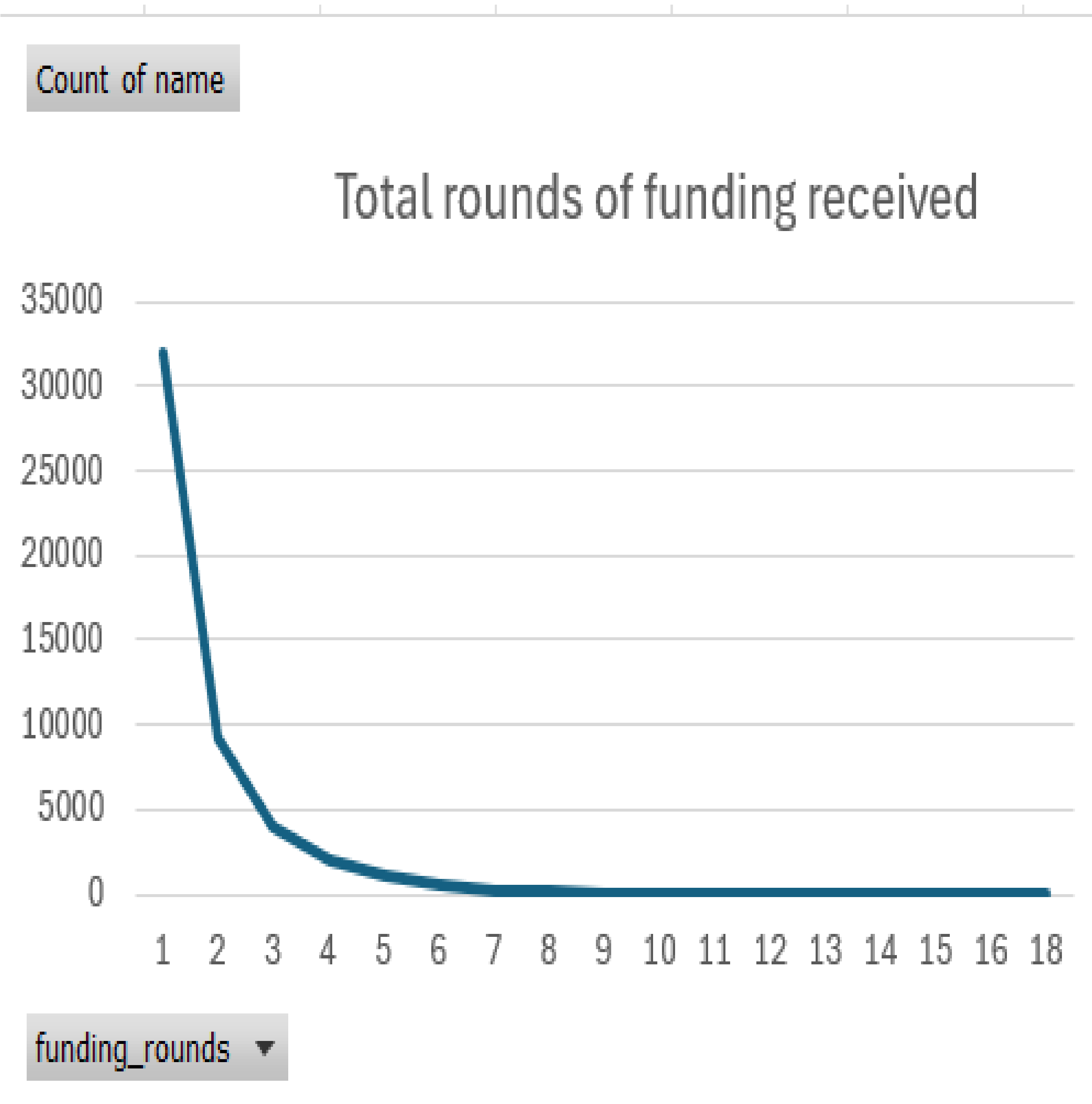

Fig. 4: *Decline curve showcasing the relationship between the total number of funding rounds completed and the volume of persisting startups.*

The right-skewed distribution of venture funding is seen in the line chart. Some 35,000 startups secure their first round of seed or early-stage funding. This number drops significantly to around 25,000 in round 2, plunges to 15,000 in round 4, and drops below 5,000 in round 7. No startup in this sample secured more than 8 rounds. That rapid dive of the first round of investments emphasizes the extremely high stakes involved in the startup lifecycle: Only a handful of companies maintain enough operational traction (that stage where capital starts moving in) to be able to unleash capital from it later. Technology is increasingly heavily competitive, and so do highly scalable sectors like SaaS, fintech, AI and the like that are generally on top with the highest early-stage funding rounds, but that drop dramatically after Series B as competition and capital intensification increase.

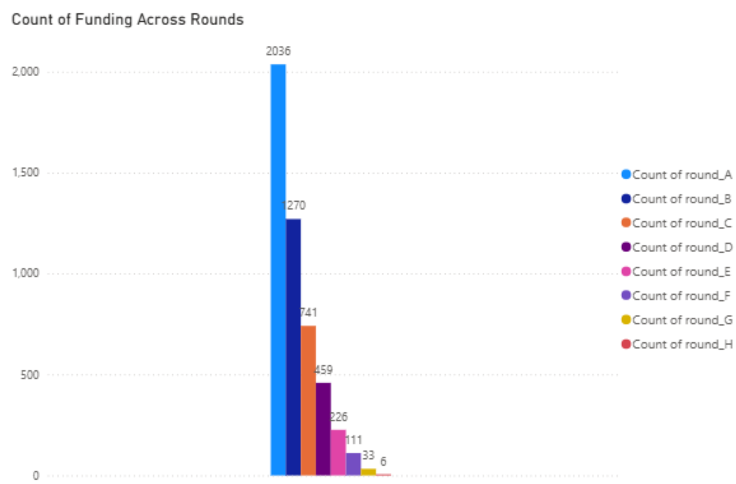

Fig. 5: *Attrition funnel demonstrating the sharp drop-off of unique startup entities advancing from Series A through late-stage institutional rounds.*

This financing funnel is characterized in the bar chart by charting specific institutional rounds. With over 2,000 startups, the sample size drops very quickly across different stages; the initial Series A round takes over the chart. Over 99% of firms that enter early stages fail to reach late-stage financing rounds (Series F through H). Moving past Series B or C in this industry takes objective market validation, clear product-market fit, and disciplined burn rates. Companies that don’t reach these milestones will promptly disappear, facing commercial or financial risks.

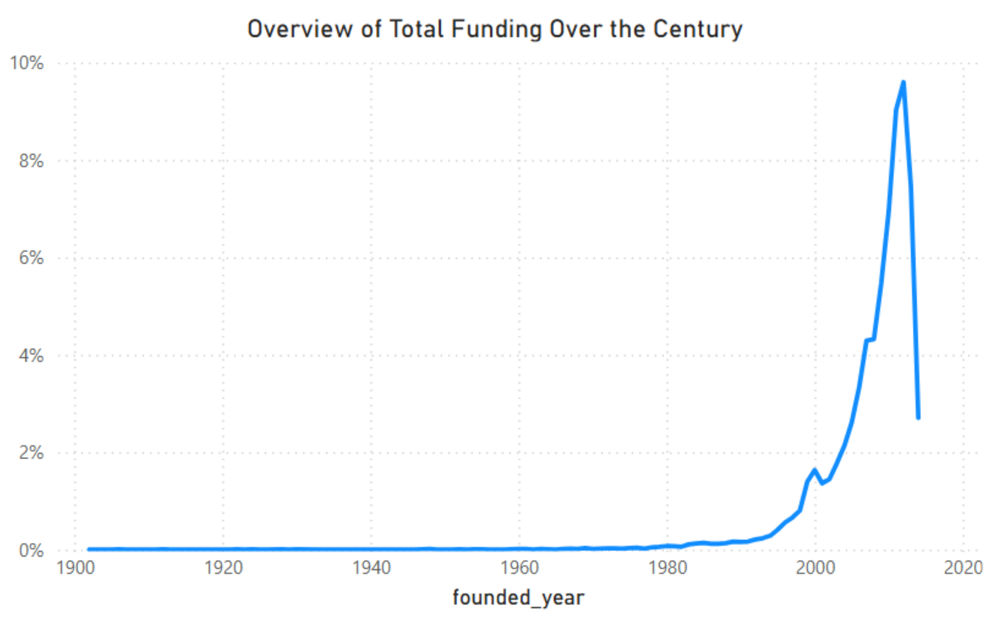

Fig. 6: *Historical trend line mapping the exponential surge in global venture capital funding allocations from 1902 through the modern era.*

A line graph illustrates the historical development of global venture financing over time. Between 1920 and 1990, the funding line remains relatively flat near 0%, reflecting an era dominated by bootstrapping, informal angel networks, and structural banking barriers. A key inflection point is reached in the mid-1990s with the dot-com boom and rise of commercial internet technologies. After a brief correction, funding volumes rose rapidly post-2000 due to low interest rates, globalized capital networks, and mega-rounds in cloud computing and mobile technology. This surge created over $2 trillion in global venture capital dry powder by 2025. Scalable software models drive over 70% of this post-2010 acceleration, supplemented by cleantech expansions and biotechnology breakthroughs.

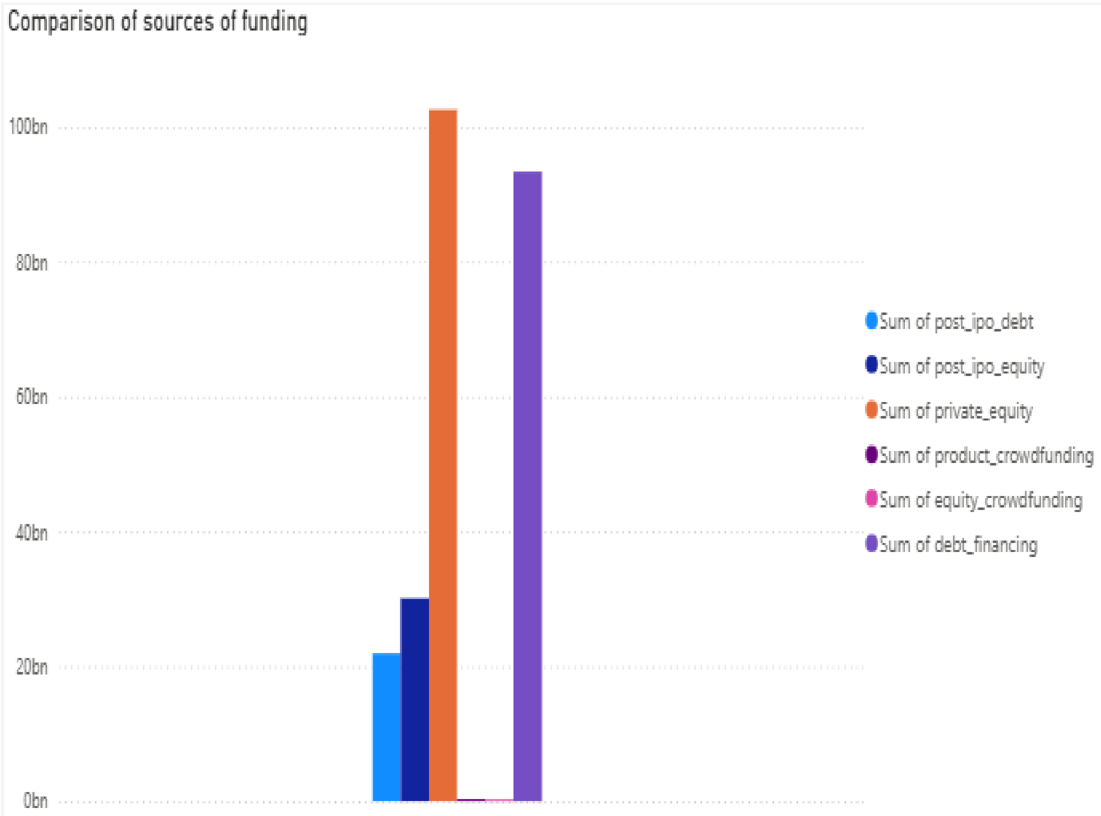

Fig. 7: *Aggregate comparison of alternative startup financing streams showcasing the dominance of late-stage private equity and corporate debt.*

The bar chart compared cumulative capital volumes of alternative funding sources. Private Equity (PE) is the main equity-backed funding source, with more than $100 billion concentrated on late-stage buyouts and growth-stage scaling. Traditional debt financing is the second largest in total with almost $95 billion, underscoring the importance of both institutional credit and corporate loans to the market, as well as traditional equity. For post-IPO equity ($28 billion) and post-IPO debt ($22 billion), on the other hand, they are small volumes, thereby suggesting that only a small number of mature startups manage to cross over to the public capital markets. And finally, alternative channels such as product and equity crowdfunding also play a negligible role, indicating that despite a major level of media attention they are still relatively minor compared to institutional private capital.

5.2. Multicollinearity and Feature Correlation Analysis

Before training our predictive models, we assessed the correlation matrix of the dataset for multicollinearity, which can affect regular regression weights' stability.

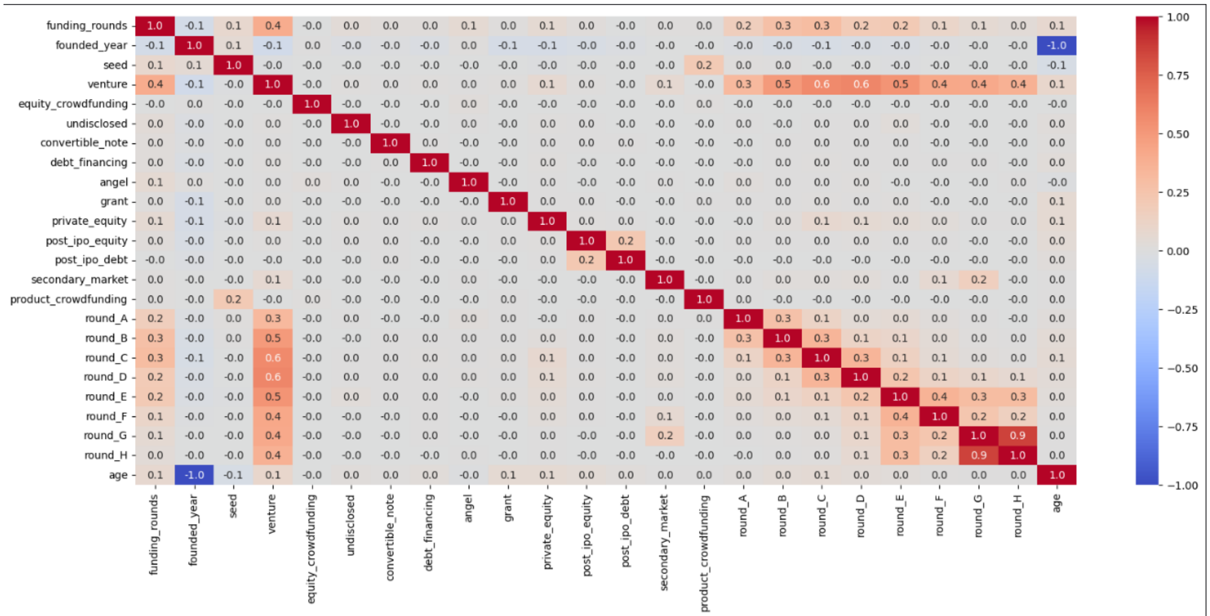

Fig. 8: *Pearson correlation matrix heatmap evaluated across independent financial, temporal, and continuous funding attributes to detect asset multicollinearity.*

The top-down correlation heatmap shows the following main structure: the firm operational age and founding year exhibit a perfect inverse correlation (-1.0). This is mathematically correct, since a firm's total lifespan is directly a function of its founding year. The number of funding rounds and institutional venture capital metrics (0.4) is highly correlated, suggesting that formal VC funding plays a significant role in the firms getting to the next phase. Round G and Round H stages were almost completely correlated over RCT time (0.9), which indicates that startups who survived to a seventh round also nearly always completed an eighth one. The lower correlation numbers for other funding sources (including government grants, angel investment, debt and crowdfunding) may indicate that these sources are not included in the traditional VC system.

5.3. Geographic and Ecosystem Clustering

The dataset shows startup activity is highly concentrated among a few geographic zones.

Fig. 9: *Global spatial distribution map highlighting the geographic concentration and localized clustering of startup ecosystems across 115 nations.*

All the recorded startup activity is organized within about 20 key countries – meaning more than 180 nations—nearly 90% of the world—show little to no presence in this dataset. Startup activity tends to be concentrated globally with strong representation in the United States, China, India, the United Kingdom, France, and Brazil. Although this clustering is to some extent due to data tracking coverage bias, it is indicative of the structural importance of localized ecosystems. High-performing startups are packed tightly around large urban centers and hubs of innovation. These high-density urban settings enable early-stage startups to secure ample labor pools, to be in close proximity to research universities, and to have established local-investor networks.

5.4. Machine Learning Model Performance Benchmarks

We compared performance on the out-of-sample 20% test slice for each of the four classification models. The empirical results are summarized and compared below:

TABLE II
Comparative Performance Metrics Across Evaluated Predictive Classification Architectures

Classification Model	Classification Accuracy (%)	Out-of-Sample ROC-AUC Score
Logistic Regression (Baseline)	71.4%	0.68
Corporate Multilayer Perceptron (MLP)	82.1%	0.79
Extreme Gradient Boosting (XGBoost)	85.3%	0.84
CatBoost Primary Classifier	**89.6%**	**0.91**

The baseline Logistic Regression model (71.4% accuracy, ROC-AUC 0.68) did not provide much performance because of its linear assumptions as well as its difficulty capturing high-cardinality categorical features. The Multilayer Perceptron performed significantly better (82.1% accuracy, 0.79 ROC-AUC) by capturing non-linear latent structures, but with a need for extensive feature tuning. XGBoost was able to achieve impressive performance, with an accuracy rate of 85.3% and an ROC-AUC of 0.84. A classifier with the best performance in all evaluation metrics was CatBoost. It achieved a classification accuracy of 89.6% and also achieved an ROC-AUC score of 0.91. For the very high-cardinality parts of a dataset, CatBoost's symmetric tree structures and native categorical processing allowed it to provide an efficient answer without increasing data dimensionality. It enabled the model to identify complex, non-linear feature interactions while retaining strong interpretability and robustness against class imbalances.

6. DISCUSSION AND FEATURE IMPORTANCE

Investigating the optimized CatBoost model, empirical feature importance rankings were prepared to reveal the key reasons that contribute toward the survival of a startup.

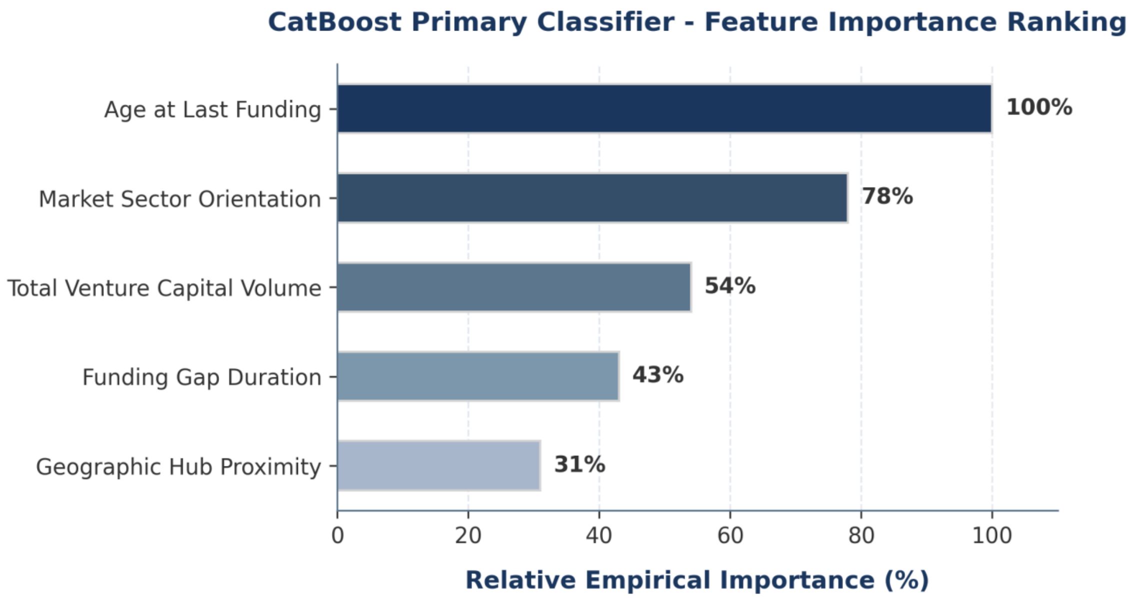

Fig. 10: *Horizontal bar chart of relative empirical feature importance ranking extracted from the optimized primary CatBoost architecture.*

6.1. Age at Last Funding

Age at last funding emerged as the most powerful determinant of long-term startup success. This measure is defined as the total time between having a company and its last funding event. The more capital the startups are consistently soliciting further along in the life cycle, the higher the survival and acquisition rates. The result indicates that an investment at a late stage is a significant factor of organizational maturity and investor confidence.. Companies that can withstand the initial market introduction and maintain their momentum in fundraising are the ones that should be able to establish a viable operation for the long term.

6.2. Market Sector Orientation

Market sector orientation of startups is second among the most important factors for the likelihood of success. These models show a clear divide between success rates in distinct verticals within their industry: Startups that operate in software, e-commerce, mobile applications, fintech, healthcare, and biotechnology are generally more likely to succeed. They have the advantage — based on the technology-driven sectors we discussed above — of high scalability, low marginal cost, and rapid market expansion, enabling firms to ramp up revenues more quickly than their total operating cost. They also closely match institutional investor interests in digital transformation, easing access to follow-on capital and acquisition exits. On the other hand, ventures in capital-intensive or low-scalability sectors like biofuels, rural energy infrastructure, custom retail, and performing arts have lower base success rates. These sectors also require large initial capital investment and a long operating time that's linked with less predictable financial returns, and less appeal to private venture capital.

6.3. Venture Capital Investment Volume and Round Tracking

The total VC investment amount and total number of funding rounds have positive and significant impacts for startup survival. Significant capital injections provide flexibility to scale up, build technology of their own, recruit top talent and meet competitive challenges. The financial support is not only significant as funding, but also as a signal to the market through funding from well-known institutional investors. This support not only bolsters the credibility of a startup but also can lead to positive network effects, making it more likely to attract further investment from other sources.

6.4. Funding Gap Duration

The time between funding rounds can offer useful insight into a startup's condition. Firms that raise capital at regular intervals, with relatively short gaps, tend to show steady progress and are more likely to meet key milestones.

On the other hand, longer gaps between funding rounds are often associated with a greater risk of failure. Such delays may reflect underlying challenges, including operational inefficiencies, missed targets, or a decline in investor confidence, all of which can increase the likelihood of cash constraints.

6.5. Geographic Hub Proximity

Location is an important factor in shaping a startup's chances of survival. Firms located in major innovation hubs—such as the San Francisco Bay Area, London, Boston, New York City, and Bengaluru—generally show higher success rates than those in less developed regions. These environments provide several advantages, including better access to venture capital, well-established accelerator and incubator programs, experienced mentors, and links to leading research institutions. While being based in such a hub does not guarantee success, the networks and resources available in these ecosystems can play a significant role in helping startups manage early challenges and scale more effectively.

8. LIMITATIONS

- This study provides valuable insights which are cycle-tested; however, there are certain limitations in the time period and scope of the data to be noted:
- **Data Cutoff at 2015:** The advantage of this machine learning model is that the dataset is deliberately limited to 2015 to maintain academic rigor and generalizability. Excluding the period 2015-2025 removes skews caused by unprecedented low interest rates, high valuations by the private sector (as seen in the heights of WeWork or Byju's), and macro-economic oddities like pandemic-induced distortions. A strict 2015 cutoff allows the model to focus on the organic factors that make things successful and not on the passing trends of the market.
- **Mitigation of Hindsight Overfitting:** Limiting the time frame prevents the model from overfitting the market or making odd predictions in the context of hindsight bias (e.g., forecasting special resilience during extraordinary global supply shocks).
- **Technical Dataset Constraints:** The time frame of this analysis was constrained by the secondary data that was available in the Kaggle/Crunchbase release, and restricted to the last_funding_at metric which is systematically ending on January 1, 2015. It was therefore not mathematically possible to continue predicting beyond this point, based on the historical setup of the raw data. Data has internal tracking coverage bias as observed in the exploratory phase (Geographic and Sector Bias). The records are predominantly in the United States and a handful of 20 countries, so the predictive insights may not be ideally applicable to startup ecosystems in unrepresented countries.
- **Geographic and Sector Bias:** Data has internal tracking coverage bias as observed in the exploratory phase (Geographic and Sector Bias). The records are predominantly in the United States and a handful of 20 countries, so the predictive insights may not be ideally applicable to startup ecosystems in unrepresented countries.
- **Survivorship Bias:** There is a potential for survivorship bias because the dataset may not include a representative number of failures that closed down before the stage that they are formally reported or institutional funding is awarded. However, the data compiled until 2015 keeps the 90% failure rate fairly realistic, which enables tree-based algorithms to generate more conservative and balanced classifications.

7. CONCLUSION

Using a longitudinal dataset of over 54,000 firms, this study examines the factors associated with startup success, with a focus on funding patterns, firm scalability, and ecosystem context rather than short-term market movements. The empirical results indicate that CatBoost performs effectively on large and imbalanced datasets, achieving an out-of-sample ROC-AUC of 0.91 and an accuracy of 89.6%.

The feature importance results highlight *age at last funding* as the strongest predictor of long-term survival. Other relevant factors include sector orientation, total venture capital raised, the time between funding rounds, and geographic location. Taken together, these findings suggest that startup success does not depend on any single variable but is shaped by a combination of structural conditions. In particular, staged financing, market scalability, and access to supportive ecosystems appear to influence firm outcomes over time.

These findings have several implications for different stakeholders:

For Founders: The results suggest that long-term performance is more closely linked to operational consistency than to short-term valuation gains. Managing fundraising cycles carefully and maintaining continuity between funding rounds can support gradual organizational development. For more capital-intensive or less scalable ventures, establishing clear competitive advantages early—such as proprietary technologies or network effects—can improve the likelihood of securing future investment.

For Venture Capitalists: The analysis reflects the highly selective nature of venture capital markets, where only a small share of firms progress to later funding stages. This pattern is consistent with the commonly observed concentration of returns among a limited number of firms. In this context, machine learning methods, including gradient-boosted models, can complement traditional evaluation approaches by helping identify patterns and assess risk in complex datasets.

For Policymakers: The findings point to the importance of strong and well-connected innovation ecosystems. Policies focused only on direct financial support may have limited impact in isolation. Greater emphasis on strengthening institutional infrastructure, supporting higher education and research, encouraging collaboration with universities, and developing incubators and accelerators may provide a more effective foundation for startup growth and long-term sustainability.

References

1. Al-Mamary, Y. H., & Alshallaqi, M. (2022). Impact of autonomy, innovativeness, risk-taking, proactiveness, and competitive aggressiveness on students' intention to start a new venture. *Journal of Innovation & Knowledge*, *7*(4), Article 100239. https://doi.org/10.1016/j.jik.2022.100239
2. Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1925695
3. Bangdiwala, M., Mehta, Y., Agrawal, S., & Ghane, S. (2022, August). *Predicting success rate of startups using machine learning algorithms* [Paper presentation]. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Pune, India. https://doi.org/10.1109/ASIANCON55314.2022.9909235
4. Binowo, K., & Hidayanto, A. N. (2023). Discovering success factors in the pioneering stage of a digital startup. *Organizacija*, *56*(1), 3–17. https://doi.org/10.2478/orga-2023-0001
5. Challoumis, C. (2024). Navigating the intersection of capital, enterprises, and AI technology. In *XIV International Scientific Conference Proceedings* (pp. 78–107).
6. Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, *10*(8), Article 1283. https://doi.org/10.3390/math10081283
7. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). *Xgboost: Extreme gradient boosting* (R package version 0.4-2) [Computer software]. CRAN. https://CRAN.R-project.org/package=xgboost
8. Drydakis, N. (2024). Artificial intelligence capital and employment prospects. *Oxford Economic Papers*, *76*(4), 901–919. https://doi.org/10.1093/oep/gpad035
9. Fiet, J. O. (2022). Risk avoidance strategies in venture capital markets. In *Venture capital* (pp. 219–242). Routledge.
10. Gompers, P. A. (2022). Optimal investment, monitoring, and the staging of venture capital. In *Venture capital* (pp. 285–313). Routledge.
11. Gourio, F. (2012). Disaster risk and business cycles. *American Economic Review*, *102*(6), 2734–2766. https://doi.org/10.1257/aer.102.6.2734
12. Kim, J., Kim, H., & Geum, Y. (2023). How to succeed in the market? Predicting startup success using a machine learning approach. *Technological Forecasting and Social Change*, *193*, Article 122614. https://doi.org/10.1016/j.techfore.2023.122614
13. Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, *15*(8), Article 255. https://doi.org/10.3390/fi15080255
14. Kritikos, A. S. (2024). Entrepreneurs and their impact on jobs and economic growth. *IZA World of Labor*, Article 35. https://doi.org/10.15185/izawol.35.v2
15. Lerner, J. (2022). The syndication of venture capital investments. In *Venture capital* (pp. 207–218). Routledge.
16. Lizarelli, F. L., Torres, A. F., Antony, J., Ribeiro, R., Salentijn, W., Fernandes, M. M., & Campos, A. T. (2022). Critical success factors and challenges for Lean Startup: A systematic literature review. *The TQM Journal*, *34*(3), 534–551. https://doi.org/10.1108/TQM-05-2021-0138
17. Ruhnka, J. C., Feldman, H. D., & Dean, T. J. (2022). The "living dead" phenomenon in venture capital investments. In *Venture capital* (pp. 349–367). Routledge.
18. Sahlman, W. A. (2022). The structure and governance of venture-capital organizations. In *Venture capital* (pp. 3–51). Routledge.

19. Sevilla-Bernardo, J., Sanchez-Robles, B., & Herrador-Alcaide, T. C. (2022). Success factors of startups in research literature within the entrepreneurial ecosystem. *Administrative Sciences*, *12*(3), Article 102. https://doi.org/10.3390/admsci12030102
20. Singh, S., & Mungila Hillemane, B. S. (2023). Sources of finance for tech startups over its lifecycle: What determines their approach of sources and its success? *International Journal of Emerging Markets*, *18*(8), 1766–1787. https://doi.org/10.1108/IJOEM-12-2021-1871
21. Skrynkovskyy, R., & Tyrkalo, Y. (2021). Entrepreneurial risks: Nature, types, assessment methods and ways to reduce them. *Path of Science*, *7*(12), 2015–2023. https://doi.org/10.22178/pos.77-7
22. Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6), Article 15501329221106935. https://doi.org/10.1177/15501329221106935