MY 498: Capstone Project

Navigating the Funding Landscape: The Role of Competition, Network Dynamics, and Patents in Fundraising Success of Indian Startups

Candidate ID: 28017

Supervisor: Dr Raphael Susewind

August 15, 2024

MSc Applied Social Data Science

Department of Methodology

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Acknowledgment

I extend my heartfelt gratitude to the faculty and academic staff at the Department of Methodology, London School of Economics and Political Science (LSE), for their invaluable support throughout my master's journey. A special thank you to my supervisor, Dr Raphael Susewind, whose expertise and guidance were instrumental in shaping both the direction and success of my dissertation. I am grateful to Crunchbase and IPqwery for providing the essential data that facilitated this study. My appreciation also extends to all my peers in this programme, whose intellectual stimulation greatly enriched my academic and research experience.

Table of Contents

1. Abstract

Despite India's rapid ascent to the world's third-largest startup ecosystem, the success of any startup industry still critically hinges on securing funding, with a significant majority failing due to financial shortfalls. Every funding milestone comes with its unique challenges, with each demanding a unique set of legitimacy signals and credibility factors that must evolve with the growth standards and the outlook of different actors in the investment landscape. This thesis explores the multifaceted dynamics of startup financing, focusing on investigating the marginal effects of competitive pressures, investor's network characteristics, and IP assets on the fundraising outcomes of Indian startups across different funding stages. Drawing from a comprehensive dataset of 8944 Indian startups founded between 2013-Jun'24, from Crunchbase, the study employs mixed-methods (NLP and Network Analysis) and longitudinal analysis approach that integrates time-series data to develop a comprehensive feature set of exogenous and endogenous factors for the multi-stage predictive model. The study reveals that investor's network characteristics significantly influences fundraising success across all funding stages, with competitive pressures lending an ambiguous effect on model predictions, owing to the duality of its impact on growth perception and crowding-out of funding opportunities. While IP assets were found to be marginally insignificant. Other key factors for fundraising success also included the number of founders, the time since inception and last funding, total funding raised, specific industry sectors, and locations. These findings can potentially contribute to the strategic decision-making processes for entrepreneurs and investors alike and offer policy recommendations to support the evolving landscape of entrepreneurial innovation and competition in India.

2. Introduction

It is estimated that 70% of startups fail in the initial two years of their operation (Failory), which underscores the need for entrepreneurs and investors to identify and understand the factors that drive startup success. A crucial aspect of startup success is the ability to consistently secure funding at each stage of company's growth trajectory and organizational lifecycle, as lack of funding or running out of funds is found to be one of the primary reasons for startup failure (CB Insights, 2023). However, every funding stage comes with its own unique dynamics and challenges, which demands that the nature and degree of legitimacy signals, that ensues confidence in a venture's potential, must evolve with the outlook of different actors (incubators/accelerators, angel investors, venture capitalists, corporate VCs) and the standard industry growth metrics, that prevail at different stages of a startup's lifecycle (Fisher et al., 2016). As a result, the factors that ensure fundraising success also vary significantly across different stages of a startup's growth. For instance, early-stage funding decisions (Pre-Series A) are fraught with higher degree of risk and information asymmetry, due to lack of proven track record and market credibility, and therefore factors that make early-stage ventures successful in attaining funding may not be equally effective during later-stages of startup funding (Islam, et al., 2018). Previously, Gastaud et al. (2019) and Stahl (2021) have studied the varying effects of competitive pressures and investor's network centrality on funding success across different funding stages. Their study suggested a multi-stage approach, sub-setting the dataset into different funding stages and training separate models for each stage to better highlight the marginal effect of these features and avoid their dilution in a global model. This study aims to develop a comprehensive and time-aware feature set that captures the snapshot of time-series factors at different points in time and explore the marginal impact of these features across different funding stages (up to Series B), through multi-stage predictive modelling. The study aims to build on the existing literature by incorporating IP related features (trademarks and patents) to the multidimensional feature set. Therefore, the research seeks to provide nuanced insights that can inform strategic decision-making for entrepreneurs and investors and offer policy recommendations to support the evolving landscape of entrepreneurial innovation in India.

2.1 Startup financing and the role of investors

Startups are seen as a critical driver of innovation, employment and economic growth in the modern economy, founded with an aim to provide solutions to either address existing gaps in the traditional markets

or create new ones. (Stahl, 2021). However, the success of startups largely depends on their ability to secure funding for growth and expansion. Raising capital is crucial for entrepreneurs to transform innovative ideas into viable businesses through the commercialization process. Investments from government grants, accelerators, angel investors, and venture capitalists (VCs) are pivotal in propelling a startup's development, allowing it to scale operations, innovate, and penetrate new markets (Ang & Saghafian, 2020). Investment amounts and equity valuations are typically aligned with the startup's funding stage—early-stage, growth-stage, and late-stage.

In the early stages, including Pre-seed and Seed rounds, investments are smaller and valuations lower, reflecting the high uncertainty and risk from the lack of proven track records. These funds are crucial for developing a minimum viable product (MVP), validating product-market fit, and achieving early traction. However, many startups deplete their seed capital before reaching key milestones necessary for further investment (Ang & Saghafian, 2020).

Successful progression past Seed rounds leads to Series A funding, where startups must demonstrate a viable business model and sufficient market traction to attract further investment. Subsequent rounds (B, C, D, E, and beyond) are marked by higher valuations and larger investments, reflecting growing revenue and market expansion. Growth-stage funding focuses on scaling operations, product diversification, and expanding market reach, while late-stage funding supports further market expansion, acquisitions, and potential IPOs.

2.2 Funding landscape in the Indian startup ecosystem

In the Indian context, the startup ecosystem has witnessed a significant surge in funding activities in the recent years, with over 1,500 funding deals realized in 2021, amounting to $36+ billion in investments, across various sectors such as e-commerce, fintech, healthtech, edtech, etc. This unprecedented growth can be attributed to a combination of factors, including the availability of global capital, a growing pool of skilled talent, supportive government policies, and an increase in internet penetration and digital awareness across the population. However, the fundraising landscape for startups in India remains highly competitive, with startups running out of funds or failing to secure funding still being one of the prominent factors for a startup's failure in India (Goswami et. al., 2023). Ghosh (2021) in his empirical analysis of factors influencing startup funding in India, found that startups in their early-stages of funding (Pre-Series A) received 40-100% lower funding, when compared to the later stages, indicating a significant gap and disparity in the access to funds at a crucial stage of startup’s growth. Likewise, even the investors face challenges in ascertaining the potential of startups, due to asymmetry of information and the high-risk nature of investments in early-stage startups. Thus, highlighting the growing need to determine the factors that influence fundraising success for Indian startups, across different stages of funding, as at each stage the investors could be looking for different signals of growth or factors of interest, to evaluate the startup's potential and make their investment decisions. (Gastaud et. al., 2019).

2.3 Motivation and Research Questions

While predicting startup success is critically significant for investors and entrepreneurs, as it offers the potential to eliminate information asymmetry and reduce the uncertainty that plagues early-stage investment decisions, it is equally crucial to understand the factors that drive fundraising success. Literature shows that funding success depends on various factors including the quality of the founding team (Żbikowski & Antosiuk, 2021), industry characteristics (Kim et al., 2023), media exposure (Sharchilev et al., 2018), competitive dynamics (Gastaud et al., 2019), and the investor's network centrality (Gastaud et al., 2019). Although recent studies have leveraged extensive databases like Crunchbase to apply advanced machine learning techniques for improving prediction accuracy (Zbikowski & Antosiuk, 2021; Ang & Saghafian, 2020; Bangdiwala et al., 2022), they often overlook the impact of these factors across different startup growth stages. Gastaud et al. (2019) addressed this by examining how competitive pressures and network centrality influence funding success, finding that competitive pressures are crucial in early stages, while

network centrality plays a greater role in later stages. This highlights the need for models that consider variable impacts at different stages of startup development.

Hence, this thesis aims to develop a robust feature set with substantial explanatory power for predicting fundraising success among Indian startups, focusing on the multidimensional drivers of such success and their varying impacts across different funding stages. The study will incorporate a blend of endogenous and exogenous factors identified by existing literature, with a special focus on understanding the marginal effects of competitive pressures, investor network centrality, and intellectual property rights on fundraising outcomes, revealing distinct patterns between early and growth-stage startups.

The research addresses three critical gaps: Firstly, the differential impact of these factors across funding stages remains underexplored. Secondly, it seeks to eliminate look-ahead bias by ensuring that the time-series variables are time-aware, reflecting only the information available at the time of decision-making. Thirdly, this study is distinctively set within the Indian startup ecosystem, which is characterized by its diverse consumer base, regulatory environment, and rapidly growing technology sector—factors that may influence fundraising success differently compared to other regions. With this context, the research intends to address the following research questions:

1) How can we quantify competitive pressures, investor's network centrality, and startup success for the predictive model?
2) What are the significant factors contributing to fundraising success of Indian startups?
3) What's the differential impact of competition, investor network characteristics, and IP assets on the fundraising success of Indian startups across different funding stages?

3. Discussion of Related Works

3.1 Features indicating funding success in data-driven prediction models

The sudden influx of literature predicting startup success through data-driven prediction models has led to the identification of several key features that can be indicative of a startup's potential to secure funding. The literature has broadly viewed these features from endogenous and exogenous perspectives (Shi et al., 2023), with endogenous perspectives being focused on factors that are intrinsic to a venture and more in-tune with the resource-based theories, suggesting inadequacies in operational inefficiencies, human and social capital, as rendering startups disadvantaged at scaling and adapting to challenges (Fern et al., 2012). Zibikowski & Antosiuk (2021), placed a significant focus on endogenous variables, such as educational background and gender composition of the founders, with a conscious aim to restrict features to only include information that known at the beginning of the company’s operation, to reduce bias against younger organizations and avoid introducing look-ahead bias into the model. Similarly, Krishna et al. (2016) included features, such as burn rate and severity scores, in their model for predicting startup success, which are indicative of the startup's ability to manage its financial resources efficiently and adapt to changing market conditions. Sharchilev et al. (2018) too aimed to include an aspect of social capital by incorporating web mentions from news articles and blogs, in addition to the traditionally intuitive features from the Crunchbase dataset. Their study found that while individual web mentions might not be impactful, their aggregation provides a comprehensive view of a startup’s public perception, thereby improving prediction quality.

Exogenous perspectives, conversely, are focused on the external environmental factors, with the potential to influence all applicable startups indiscriminately. These typically manifest in forms of competitive pressures, regulatory environment, and industry and market dynamics, and macroeconomic conditions. For instance, Kim et al. (2023) leveraged Crunchbase data to include nuanced industry characteristics, such as novelty of industry, industry persistence and centrality measures of industry groups from industry co-classification network, with an aim to control for the impact of entry barriers, market segmentation, industry popularity, etc. that potentially results in some industry/sectors being linked with higher survival rates than others. The

study found the centrality measures to be significant in predicting startup success, suggesting that firms in attractive and highly convergent industries are more likely to succeed, as it also coincides with new product development and technology convergence across such industries. However, the reliability and validity of the industry features from this study can be questionable, considering the industry classification tags on Crunchbase are selected by the business owners, and hence could be misleading or motivated by different agendas. Further, macroeconomic factors highlight that negative economic shocks, capital market fluctuations, and restrictive regulatory policies increase the probability of startup failure by limiting access to financing and growth opportunities (Donskikh, 2021; Kuckertz et al., 2020).

According to network theories, startups positioned on the periphery of business ecosystems struggle to access vital resources, knowledge, and partnerships, and deprives them of critical support mechanisms (Soda et al., 2004). To demonstrate the network theory in predictive context, Bonaventura et. al.(2020), utilized time-varying startup networks to predict startup's future success, wherein each edge between startups represents a shared individual (employees, board members, investors, founders). Their findings reinforced the theory that the position of a start-up within its ecosystem is relevant for its future success. This position is highlighted by factors such as successful hiring, media attention, and the backing of influential investors. Furthermore, in a unique departure from the traditional approach, Gastaud et al. (2019), focused on exploring the marginal effect of competitive pressures and investor's network centrality on fundraising success, finding that competitive pressures are more influential in early-stage funding rounds, while investor's network centrality becomes more prominent in later stages. Therefore, this study aims to integrate a variety of both endogenous and exogenous features into the predictive model, to provide a more comprehensive theoretical lens on the multidimensional drivers of startup success in the real world (Shi et al., 2023).

3.2 Definition and quantification of success labels

What constitutes success for a startup? This question that has long been debated in the literature, with various studies using different metrics to define and quantify success, owing to its non-definitive and subjective nature. Kim, et. al (2023) considered IPO as the sole proxy for success, and consequently ran into the challenge of a significantly imbalanced dataset, as only a small fraction of startups ever reaches the stage of an IPO. Xiang, Zheng et al. (2021) identified a startup being acquired as an indicator of success. However, this measure is not ideal due to its sparsity and unclear outcomes. Krishna et. al (2016), Singhal et. al (2019) on the other hand, used a more broader definition of startup success, focusing loosely only on the survival of the company, by considering all metrics: IPO, M&A and any subsequent funding round(s) between inception and prediction instance, as the indicators of success. This is a poor prediction target as being in business may not necessarily assure success for the investors.

According to Sharchilev et. al.(2018), the definition of success must satisfy two main conditions: First, it should translate to real profitability. Second, it should be available for evaluation at the prediction instance and should not require forecasting into the distant future, to maintain tractability. For entrepreneurs, success could be defined by non-financial metrics, such as the ability to create a sustainable business model or achieve higher customer satisfaction. However, for investors, success has invariably been equated with maximizing financial gains and the return on investment (ROI). This increase in ROI is generally achieved when the market valuation of the startup increases, thus pushing up the notional value of the investor’s equity share in the company. With this context, it can be argued that IPO or merger & acquisition (M&A) events are the only ultimate paths to success, as they provide the investors with an exit opportunity to realize their returns on investment. However, both these events can be an imperfect proxy metric for success. That's because relatively few startups ever reach to the stage of an IPO, and if they do, it is quite late in their development trajectory. Therefore, due to these lengthy incubation periods between the inception of a startup and its eventual exit, labelling a startup that is progressing towards success but has not yet fully achieved it as a failure would be an unfair evaluation (Ang & Saghafian, 2020). Likewise, using mergers and acquisitions (M&A) events as a proxy for success is flawed in terms of accuracy and coverage, since not all

successful companies undergo M&A and, crucially, only a subset of acquired companies go on to be successful and deliver financial gains to their shareholders. (Moeller et. al., 2005)

Moreover, Stahl (2021) posits, that startup's ability to raise subsequent rounds of funding at higher valuations amounts to a better measure of success, as it not just increases the post-money valuation but also provides early-stage investors with the opportunity to cash out their investments through secondary transactions to later-stage investors in the follow-on rounds. Sharchilev et al. (2018), established that securing funding serves as a strong indicator of a startup's present or potential business value, validated by an investor's confidence in the company's potential, who can be considered a knowledgeable expert in the field. A notable advantage of predicting the types of funding rounds a startup might attract is the flexibility this method offers: by adjusting the target round, one can balance the level of risk against the potential reward desired by an investor.

4. Research framework

4.1 Data collection and pre-processing

The primary data for this study has been sourced from *Crunchbase*, that is widely regarded as the world's most comprehensive open data set about startup companies (Retterath and Braun, 2020). We used a uniquely generated REST API user-key, made available as part of the platform's Academic Research Access program. The data was extracted using the organization, entities, and people endpoints from the Crunchbase API, that allowed the use of custom search queries to filter and retrieve a list of all Indian startups founded between 2013-Jun'24. The dataset comprises of 8961 companies with information on their location, textual description, category keywords, industry, founding date, operational status, etc. The dataset was further enriched by querying individual company profiles using their unique identifiers, to extract company specific information, such as funding rounds, investors, founders, and media reference related information.

To achieve higher data completeness, automated and manual efforts were attempted to extract the missing descriptions for the 382 startups, from their LinkedIn profiles, using HTML scraping techniques. The efforts resulted in successful extraction of 257 of the missing descriptions. The remaining 125 observations were deleted, due to downstream dependencies on the description variable for identifying competitors. Further, the data was cleaned and pre-processed to remove duplicates, extract the relevant datapoints from the nested lists, and convert the dates to a standard date format.

The data was further augmented with information on patents filed by the startups, sourced from PATENTSCOPE database, powered by WIPO (World Intellectual Property Organization). The patent data was extracted using an automated Selenium driver, and by querying the PATENTSCOPE online database with the legal names of the startups, required to match the Applicant Name of the patent. To ensure accuracy of results, the filters were applied to return only patents applicable to the Indian jurisdiction, and some manual screening of results was performed. The results returned counts of applicable patents for 573 companies. The patents data was further augmented with data on registered trademarks, obtained from *IPqwery*, an aggregated database of intellectual property profiles of fortune 500 companies and global startups.

4.2 Feature engineering

The prior discussion of endogenous and exogenous perspectives on factor identification, offers a strong theoretical framework to guide the selection of key features, having the potential to explain the fundraising success of startups. For instance, the resource-based theory suggests the importance of capturing intrinsic variables, relating to founders, intellectual and technology assets, capitalization etc. Similarly, positioning and experience considerations point towards the potential of including factors, such as location, startup age, industry, etc. While the exogenous perspective underscores the importance of controlling for temporal and environmental indicators, such as market competition, macro-economic conditions, investor's network

dynamics etc. This way, the theoretical underpinnings guide feature engineering for this study, and allow to empirically test their validity and hypothesized predictive relationships with the startup outcomes. The study proposes to categorize the features into six feature groups: General features, Financials and operations, Competitive pressures, Investor's network and performance, Industry Characteristics, and Technological innovation.

4.2.1 General features

The general features include basic information of the startups and founder attributes, such as startup's age(in days), location of the headquarter, number of founders, previous startup experience of founders, share of female founder in the team, industry sector, etc. These features are expected to provide a foundational understanding of the startups and their business context. For instance, founders with extensive previous startup experience are likely to develop critical knowledge about navigating the key milestones from ideation to scaling, be equipped with the necessary skills to manage the startup's operations (fundraising, hiring, product-market validation, etc.), and even come with a richer network of contacts and resources that can be leveraged for the startup's growth (Shi et al., 2023). Similarly, the research also shows that startup-specific traits, such as industry sector, location and age, can offer insights into the startup's maturity and the context in which it operates, thus complementing assessments based on financials and founder attributes. (Lee et al., 2018; Moskovkin, 2020; Sengupta et al., 2023). However, given the previously highlighted challenges with reliability of the industry classification/category tags from Crunchbase, being owner generated and potentially misappropriated, the study leverages the textual descriptions of the startups to extract the industry sector, using LDA (Latent Dirichlet Allocation) based topic modeling, such as STM (Structural Topic Modelling). Topic modelling (TM) assumes each word follows a two-step process: iteratively assigns topic distributions to descriptions and word distributions to topics, refining until the best model is found. Unlike simple keyword counting, TM considers word context, capturing nuanced semantic relationships (Savin et. al., 2022). Lately, a few of the contemporary studies have employed TM into areas similar to our research. For instance, Ang & Saghafian (2020) extracted 16 industry sectors (topics) using LDA, and further used them in their predictive model, only to find a highly significant impact of industry sectors on determining startup success.

Furthermore, Färber and Klein (2021) found the share of female founders to be negatively correlated to the funds raised in the early stages, which could be an implicit bias in the investment decisions, and hence it could be insightful to control for this variable in the model. Therefore, this combination of data on startup context, along with quantification of viability and growth potential, can lend more explanatory power to the model and help identifying the underlying factors that drive fundraising success.

Proposed Solution:

For text preprocessing, we utilize the 'quanteda' package to convert the descriptions into a corpus, applying standard steps such as tokenization, lowercasing, and removal of stopwords, punctuations, URLs, and symbols. A document frequency matrix (dfm) is created with a minimum term frequency of 10, then converted into an STM-compatible format. To determine the optimal number of topics (k), we consider model performance based on three criteria: Residuals (measure of how good the topics fit the data), Semantic coherence (degree of co-occurrence of most probable words in a topic), and Exclusivity (weighted harmonic mean of the word’s rank in terms of exclusivity and frequency). Based on the diagnostic outputs, we select the optimal number of topics to be k=26, that balances the trade-off between the three criteria, along with providing the most nuanced sector distinctions.

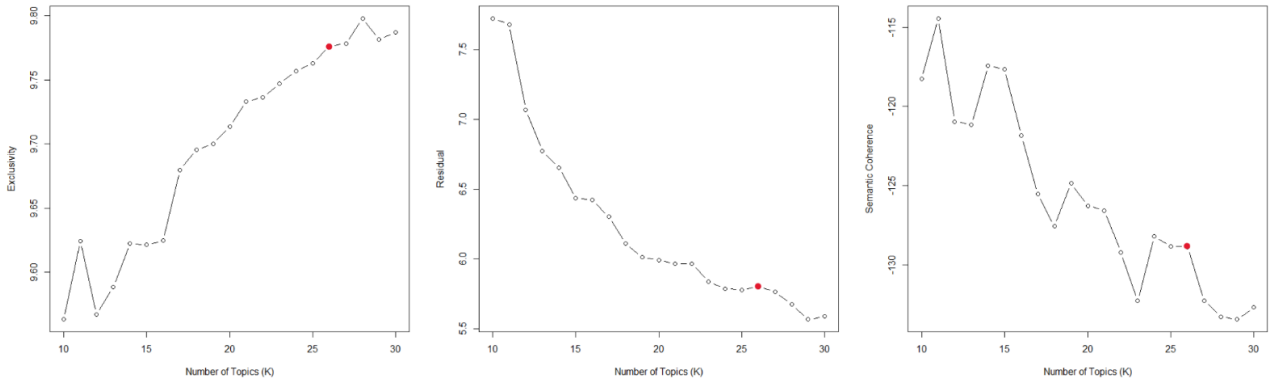


Fig. 1: Diagnostic output for Exclusivity, Residual, and Semantic Coherence.

The applied STM model generates meaningful topics and offers details on the 26 industry sectors identified based on the company descriptions, as shown in Table 1.

Topic #	Topic FREX words (n=10)	Assigned Industry sector/Topic
1	fleet, logistics, delivery, truck, transportation, transport, deliveries, express, mile, freight	Logistics & Trasportation
2	car, homes, rent, rental, repair, interior, home, furniture, parking, appliances	Home & Utility
3	estate, group, offices, ventures, real, countries, tier, years, global, east	Real Estate & Venture Capitals
4	celebration, people, friends, love, outstation, believe, always, idea, something, way,	Social Community Experiences
5	credit, loans, loan, lending, banking, fintech, finance, lenders, banks, financing	Financing & Banking Services
6	hiring, recruitment, candidates, talent, job, workers, training, jobs, recruiters, employers	Talent hiring & skilling
7	marketing, web, advertising, seo, agency, development, campaigns, digital, hosting, software	Digital Marketing & Web Development
8	contact, networking, email, conversations, discuss, calls, videos, chat, call, pages	Connectivity & Communication Solutions
9	patients, patient, diagnostic, chronic, medical, healthcare, doctors, clinic, clinical, ambulance	Healthcare
10	sports, hotel, travel, cricket, travelers, fitness, dating, trip, bus, booking	Travel, Sports & Fitness
11	esports, news, gaming, stories, games, content, languages, entertainment, readers, publishers	News, Media & Entertainment
12	web3, blockchain, building, crypto, decentralized, cryptocurrency, developers, layer, defi, exchange	Cloud & DeepTech
13	iot, industrial, drone, technologies, industries, engineering, security, robotics, drones, aerospace	Robotics & Industrial Technology
14	earn, users, influencers, allows, sellers, merchants, deals, stores, enables, offline	Online Marketplace
15	legal, corporate, event, compliance, hr, employee, payroll, management, communications, firms	EnterpriseTech & Professional Services
16	supply, farmers, chain, farming, agri, agricultural, crop, farmer, agriculture, farm	Agriculture & Supply Chain
17	insurance, funds, tax, stock, mutual, investing, claims, investment, wealth, policies	Insurance & Financial Investments
18	fashion, beauty, jewelry, apparel, clothing, accessories, hair, men, baby, skincare	Fashion, Beauty, & Retail
19	headquartered, founded, delhi, karnataka, bangalore, maharashtra, haryana, mumbai, hyderabad	Limited Information*
20	ai, intelligence, analytics, machine, data, artificial, insights, ai-powered, ml, workflows	Data Analytics & AI
21	teachers, students, exams, student, education, educational, school, courses, educators, e-learning	EdTech
22	snacks, beverage, tea, teas, flavors, tasty, food, meat, foods, beverages	Food & Beverage
23	electric, ev, charging, battery, batteries, mobility, energy, solar, motors, renewable	Electric Mobility & Renewables
24	android, ios, download, music, app, mobile, party, apps, pet, smartphone	Software & Mobile Apps
25	augmented, vr, virtual, ar, reality, university, immersive, metaverse, 3d, creative	Design & Graphics
26	consumer, end, internet, empowering, consumers, gap, aims, age, tech, solving	E-Commerce & Consumer Technology

Table 1: Topics generated from STM and their correspondingly assigned industry sectors.

The second column displays the top 10 words, identified using the FREX method, favoring frequent and exclusive words within a topic. Topic labels were manually assigned based on FREX words and a thorough review of the top 10 descriptions under that topic category. Further, based on the probability distribution of topic prevalence across each startup's descriptions, which sums to 1 across the 26 topics, we assign the industry corresponding to the topic with the highest prevalence for each startup. For instance, the description for BluSmart Mobility, an on-demand electric cab service, shows high topic prevalence of approximately 0.2 for Topic 1: Logistics & Transportation and 0.4 for Topic 24: Electric Mobility & Renewables. Thus, it is categorized under the Electric Mobility & Renewables industry.

**Notably, Topic 19 is labelled as 'Limited Information', since it reflects a lack of specific information about the startup's business model, industry sector, or product/service offering, and only captures the location specific details. Therefore, Topic 19 was excluded from being assigned to any of the companies, as it does not provide any meaningful insights into the startup's industry sector or business context.*

4.2.2 Financial and operational features

Funding is certainly critical for the survival of startups, which is reinforced by the finding that running out of funds or failing to raise new ones forms the second biggest reason for startup failures, accounting for 29% of startup closures (CB Insights 2023). Consequently, existing literature has also found that startups that are well-funded or VC-backed outperform their non-funded counterparts. For instance, Zava & Caselli (2023) show evidence that a higher funding amount during the seed-stage is positively correlated with the startup's ability to raise subsequent rounds of funding. Therefore, incorporating financial and operational features, such as the total funding raised, the number of funding rounds, the time since the last funding round, time taken to raise first invetsment, etc, can provide insights into a startup's financial health and business fundamentals (Shi et al., 2023). These group of features are expected to capture the startup's funding attractiveness among investors, ability to manage its financial resources efficiently and sustain its operations in the long run.

4.2.3 Competitive pressures

From an exogenous perspective, competition is a critical environmental factor that investors consider when valuing the startups. Gompers, Kaplan, et al.(2020) found in their study that 7% of early-stage investors and 11% of late-stage investors considered a startup’s market and the competition it faces as the most critical factor during deal selection. The presence of strong competitors can pose a threat to a startup's potential market share, profitability, and long-term viability, making it less attractive to investors. Therefore, a highly competitive landscape can significantly impact a startup's ability to secure funding, as investors could perceive them as high-risk and low potential investment. However, on the contrary, competition can also be seen as a sign of a growing market and a validation of the startup's value proposition. And as noted by Kaplan & Stromberg (2004), VCs cite large and growing markets as attractive in almost 69% of the investments; and a favourable competitive position and a high likelihood of customer adoption, in roughly 30% of the decisions. This suggests that with the right differentiation strategy and product innovation, startups can even leverage competition to their advantage, by capturing a relatively small, but an absolutely large share of a growing market, with an existing and validated demand. Therefore, given the duality of its perception, competition can be a critical factor to determine which way the investor's decision might sway, and yet remains a neglected feature, except for a few studies (Gastaud et al., 2019; Stahl, 2021) that have attempted to quantify and factor-in the impact of competition on startup’s ability to raise funds.

Stahl (2021) highlights the relative nature of competitive pressures, wherein competitor's behaviour can govern the action taken by a startup, and hence the startup is expected to react when its competitors start to raise more money. Therefore, for the purpose of this study we use the number of competitors and derive two further features: the average of funds raised by a startup's closest competitors within the preceding 12 months, as a proxy for competition in the funding landscape, along with the proportion of competitors who have actively raised funds in the same period, to reflect the effective and immediate competitive pressures faced by the startup. However, since there is no reliable data available on competitors of the startups, we utilize the free-text description of the startups from Crunchbase, which has found extensive use in past literature for gathering business insights using various NLP tasks such as topic modelling and semantic similarity. Shee, Lee et al., (2016) utilized the textual descriptions of US based companies from Crunchbase to measure dyadic business proximity of firms, using the LDA(Latent Dirichlet Allocation) topic modelling technique. It further applied the business proximity measure to analyse M&As in US high-tech industry. And similar to our context of use, Gastaud et al. (2019) and Stahl (2021) employed Word2Vec+SmoothInverse Frequency and BERT sentence embedding models, respectively, to identify the closest competitors of a startup, based on the similarity of their descriptions.

Leveraging from the existing literature, this study identifies the closest competitors by applying textual similarity measures, such as cosine similarity, to the sentence embeddings of startup descriptions, retrieved using the Universal Sentence Encoder (USE) model. Perone et al., (2018) compared popular sentence embedding techniques on various downstream NLP tasks, like text classification, semantic text simalirity, etc., and found USE(Transformer) to have outperformed the other models, including Word2Vec, ELMo, and InferSent, on majority tasks related to evaluating semantic similarity. Therefore, for our use case we propose to employ the Universal Sentence Encoder (USE) model, that leverages a Transformer architecture to encode sentences into high-dimensional vectors, capturing the context-aware semantic similarity between them.

Proposed Solution:

1) Generation of sentence embeddings using the Universal Sentence Encoder (USE) model.
2) Calculation of similarity scores using cosine similarity, between each pair of sentence embeddings.
3) Identification of the closest competitors based on a threshold similarity score (0.5).

The similarity between two embeddings was measured using standard cosine similarity: given two vectors A and B, the similarity between them is calculated by:

$$\text{similarity} = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

The study considers startup pairs with a cosine similarity score of 0.5 or above as the competitors of each other, which is set arbitrarily at a value that is neither too high, so as to not return any valid results, nor too low to let in irrelevant comparisons as close competitors.

Below are comparisons of the competitors identified (at 0.5 threshold) by the different sentence embedding models, retrieved for companies selected at random:

Company	Company Description	Model	Competitors [count of competitors]	Observation/remarks
Sploot	Sploot is a community-led platform for better pet parenting with the aim to create happier and healthier lives for pets by empowering pet parents with the right knowledge……..	Doc2vec (DBOW)	Petmojo , Pet Chef, Veg Route, Pawfectly Made, myHarvest Farms, FLOAP , GSM AND CO, Nutty Yogi, Furry Feedz, Babygogo, Hopping Chef, Laundrokart [12]	While the model clearly identifies a few companies in the Pets category, it also included irrelevant companies such as, Veg Route - which is an online grocery delivery service.
		USE (Transformer)	Petmojo, Snouters, PetSutra, GoPetting, TailsLife [5]	The model accurately identifies only Pet related firms, but also the theme across all the companies is similar – that of pet care/pet parenting, unlike Doc2Vec, which mostly included companies dealing in pet food and products. Thus, reflecting a highly context aware embedding of sentences by the USE model.
		BERT	Ibots.in, SetEducation, Internset, Petmojo, Karigar…..[3297]	The count of identified competitors is unreasonably high to assume it to be accurate or valid.
Datamotive.io	Datamotive is a firm that provides the seamless migration of workloads in a hybrid multi-cloud environment with a guaranteed SLA of 10 minutes. Its unique approach safeguards enterprises against data loss caused by cloud security misconfigurations, data breaches, or corporate network corruption.	Doc2vec (DBOW)	Moshak Tech, TTSF CloudOne, Nimesa Technology, BizCarta India, Lumiq [5]	The model picks out TTSF CloudOne – a cloud kitchen company, as similar toa cloud security and data migration company. Thus, reflecting the context insensitivity of the model.
		USE (Transformer)	Cloudanix, Moshak Tech, SynctacticAI, Seconize, EnCloudEn, Nxtra Data [6]	All companies identified as competitors are exclusively into cloud technology or data migration or cybersecurity, which validates the accuracy of the model.
		BERT	Finteck, Newme, Creditail, Propdata…….[5686]	The count of identified competitors is unreasonably high to assume it to be accurate or valid.

Table 2: Empirical evidence comparing the quality of competitors identified by different sentence embedding techniques.

4.2.4 Investor's Network Centrality

The role of investors in the growth of startups has been actively studied, but proportionately less attention has been paid to the network dynamics of investors and their impact on startup success. Network analysis can allow to reveal the investor's network characteristics through a syndication network, that can be represented as a graph where the nodes are all the investors in the startup ecosystem and the edges reflect the joint investments made by them in a common venture. As a practice, VCs prefer to syndicate their investments with other VCs rather than investing independently (Lerner, 1994a). This practice ties them into networks of relationships formed through their current and previous investments with other VCs. The pioneering effort to link VC networks with the growth of their portfolio companies is credited to Hochberg et al. (2007). They assert that companies backed by well-connected VCs have a significantly higher likelihood of surviving to secure additional financing and achieve a successful exit. This positive impact of investor's network structures has been further corroborated by Lim et al. (2018), who found that investors not only contribute financial capital to a venture, but also act as conduits of knowledge diffusion, technology convergence and social capital, by virtue of their position in syndication, as well as other social networks. For instance, Hadley et al. (2018), attempted to study the impact of VCs on startup success using social network analysis, such as the interlocking directorates network (linking startups with a common board member) and the Twitter social network of board members. They reported that startups with a central position in the interlocking directorates network were more successful with respect to funding and sales revenue, with socially well-connected VCs on Twitter as their board members. Although the presence of VCs was observed to negatively impact the ROI of startups, compared to Non-VCs.

And similar to our context of study, Caselli & Zava (2023), employed a discrete-time dynamic network model to develop a novel funding attraction index, capturing the temporal influence of investors on the fundraising success of startups. And observed higher likelihood of startups raising Series A funding when their early-stage investors have high attraction scores. The study also examined the transition from Series A to Series B, noting a decrease in the explanatory power of the attraction index in later funding stages. Which seemed contrary to, Gastaud et al. (2019), who found that investor's network centrality measures become more prominent at explaining fundraising success in later stages of funding. Therefore, the marginal effect of investor's network position on startup success remains uncertain, with varying results across different studies.

Leveraging the use of syndication networks from past literature, this study incorporates both the temporal and bipartite dimensions of investor-startup network relations. At the end of every six months, a bipartite network is created with investors and startups as the two types of nodes and funding events as the connecting edges. This network is then projected onto the investor nodes to form a syndication network, enabling the calculation of centrality measures such as degree centrality, eigenvector centrality, and betweenness centrality of investors. In the context of the syndication network, degree centrality captures the basic connectivity by reflecting the number of unique investors it has syndicated with. Eigenvector centrality measures the influence of an investor by assessing their closeness to other well-connected investors. Betweenness centrality identifies strategic bridging roles by calculating the number of shortest paths between nodes that an investor occupies. These measures are expected to collectively capture a comprehensive view of the investor's network characteristics, to determine their impact on startup's ability to raise funds, across different funding stages.

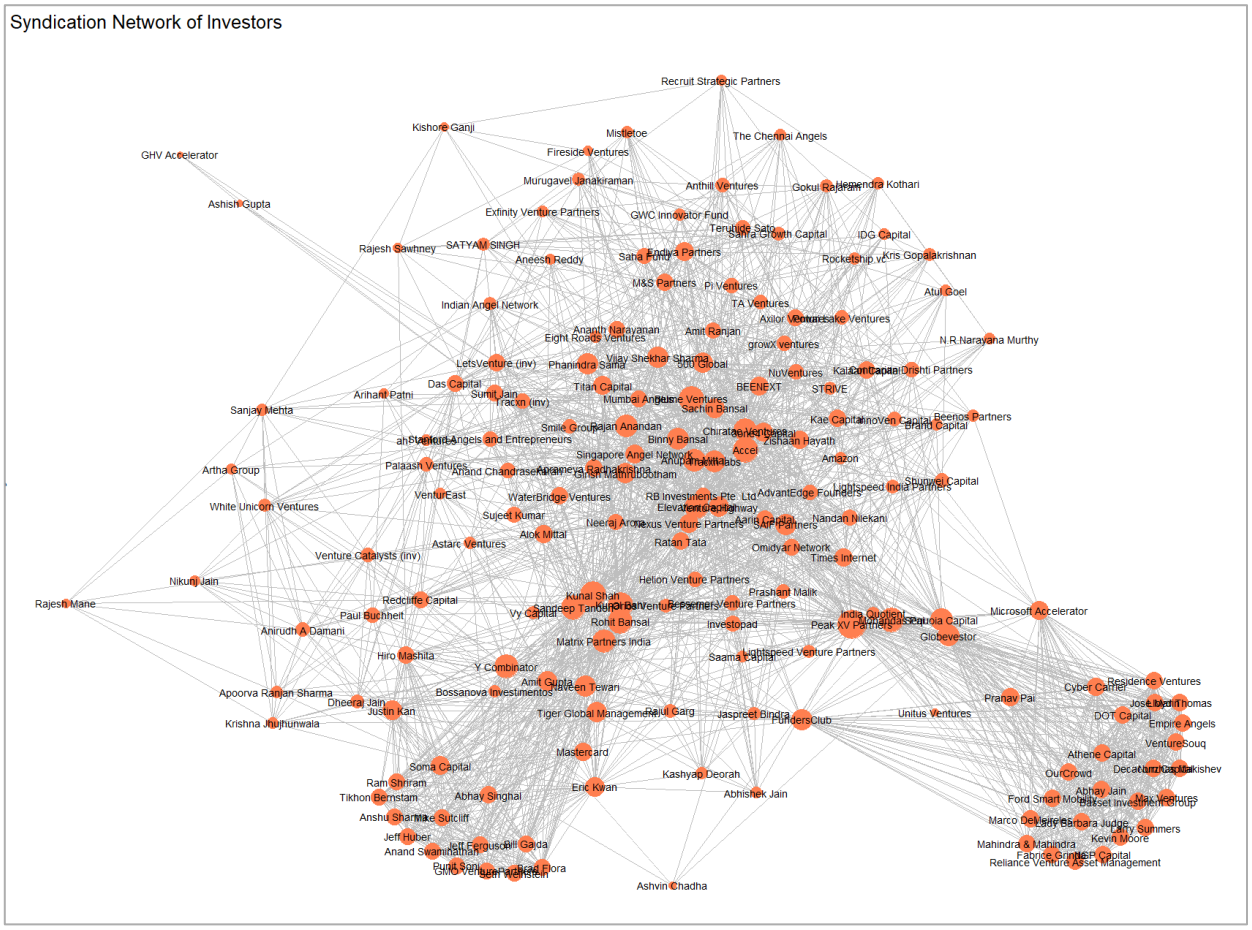


Fig. 2: Sample plot of a subgraph of one of the investor syndication networks up till Jun'2018.

4.2.5 Intellectual Property features

Intellectual property (IP) rights are considered to be at the cornerstone of innovation and technological advancement, providing startups with a competitive edge in the market. Patents, in particular, are a crucial form of IP protection that contribute significantly towards legitimizing a firm's technological capability and intellectual capital. They can thus be an effective signal to overcome informational asymmetries during early-stage funding and inform investors about the startup's innovation potential and competitive advantage. Previous studies have noted that intellectual property is imperative for startup valuation (Hsu and Ziedonis, 2013; Kohn, 2018). In particular, patents serve as a crucial indicator for seed investment, especially for technology-oriented startups (Hahn et al., 2017). However, Keogh & Johnson (2021), found that patents as standalone weren't significant at reducing the failure risk of startups, but with other pieces of IP factored in, such as copyrights and trademarks, the impact was significant. Similarly, Block et al. (2014) indicated that trademarks and patents can influence startup valuation by safeguarding against intellectual property infringement by competitors and signalling to the market that the firm possesses advanced technological capabilities. Therefore, the inclusion of IP-related features, such as the number of published patents and number of registered trademarks owned by the company, can provide valuable insights into the startup's innovation potential and its ability to secure funding.

4.2.6 Feature selection and preprocessing

The final dataset consists of 6445 observations/rows and 70 columns. The feature set can be classified into five feature groups, with a total of 22 features: General features, Financial and Operations, Competitive Pressures, Investor's Network Centrality, and Intellectual Property. The considered features are of three types: Numerical, Time-series, and Categorical. The time-series variables are all time-stamped and re-captured at specific time intervals (end of every semester) within the period of study, that coincides with

prediction instances, to incorporate the applicable updates for the latest funding rounds in the sample. For pre-processing of data, the study used the following methods: log-transformation for skewed numerical features, one-hot encoding for categorical features, and min-max scaling for standardization of numerical features. The total list of features used in the predictive model is provided in Table 3.

Feature Group	Feature	Type	Preprocessing
General	Days since founded	Time-series	Min-max scaled
	Days since last funding round	Time-series	Log transformed & Min-max scaled
	Number of founders	Numerical	Min-max scaled
	Proportion of female founders	Numerical	
	Number of previous startups founded by founders	Numerical	Min-max scaled
	Number of media references/news articles	Time-series	Log transformed & Min-max scaled
	Industry sector	Categorical	One-hot encoded
	Location of headquarters (State)	Categorical	One-hot encoded
Financial and operational (F&O)	Total funding raised (till that instance)	Time-series	Log transformed
	Funding raised in the last round	Time-series	Log transformed
	Number of previous funding rounds	Time-series	Min-max scaled
Competitive pressure (CP)	Number of close competitors	Time-series	Log transformed & Min-max scaled
	Avg. of funds raised by competitors in last 1 year	Time-series	Log transformed
	Proportion of competitors who raised funds in last 1 year	Time-series	
Investor's Network Centrality (INC)	Number of investors	Time-series	Log transformed & Min-max scaled
	Max. of betweenness Centrality	Time-series	
	Mean of Degree Centrality	Time-series	
	Max. of Eigenvector Centrality	Time-series	
Intellectual Property (IP)	Number of published (issued) patents	Time-series	Log transformed
	Number of registered trademarks	Time-series	Log transformed

Table 3. Feature groups and preprocessing: Displayed are the feature groups with associated features discussed in section 4.2. Further, it displays the type and considered pre-processing methods for each feature.

4.3 Defining success: Funding rounds

As pointed out earlier, raising funding rounds is a crucial indicator of startup success, signalling a startup's growth trajectory and competitive advantage in achieving investor-attracting milestones through various funding stages. For investors, early identification of startups likely to secure funding offers significant advantages, enabling them to tailor investment strategies and gain a competitive edge for sustainable returns in a competitive market (Stahl, 2021).Therefore, the study proposes to define success as the ability of a startup to secure funding at different funding stages, with successful startups being those that are able to raise capital, go public, or get acquired within the time horizon of next 2 years, aligning with literature that identifies funding as a fundamental determinant of startup success.

4.4 Candidate selection

Having defined success as the ability to raise funds, go public, or be acquired within the next t years, the remaining question is what to use as the time-reference point for predicting success over this t time horizon, and which observations should constitute the samples for the model. Sharchilev et al. (2018) suggested two approaches to define the time-reference point: a company-centric approach, which involves making predictions 'n' days after the last funding round of the company, and an investor-centric approach, which involves fixing a point in time and making predictions from this date for each candidate startup.

This study opts for the latter approach, where funded startup samples are accumulated across every semester (a 6-month period) between 2013 and June 2022, and the success of a startup may occur in the future relative to the applicable semester window. This approach not only closely mirrors real-world investment scenarios but also offers other advantages:

First, it enables data augmentation, meaning that each company may appear multiple times across the training and test sets, representing different funding events over the semesters, thereby increasing the sample size by an order of the magnitude of the funding rounds raised by the startups, within the study period.

Second, after splitting the data into training and test sets, a particular company's snapshots taken at different moments may appear in both sets. This is not considered data leakage because a startup's features and prediction targets evolve over time. Further, to ensure there is no look-ahead bias in the model, the study uses only features that are available at the time of decision-making (end of each semester), and do not require forecasting into the future. This is ensured by the use of mostly time-stamped features and taking a snapshot of the time-series features until that moment in time, such as the number of funding rounds, the total funding raised, the number of published patents, number of press references, etc.

Third, is the ability to utilize a multi-stage strategy for highlighting the varying impact of features across the different funding stages. As noted by Stahl (2021), most successful venture capitalists adopt a multi-stage investment strategy that spans from early to later-stage startups. However, these multi-stage strategies have received limited attention in academic research. This study introduces a multi-stage success prediction model by segmenting the dataset according to different funding stages and developing separate models for each. This method enhances the precision of the analysis by emphasizing the unique effects of features at each stage, thereby preventing their generalization across a universal model. This approach aligns with our previous observation that early-stage and growth-stage funding rounds differ in deal dynamics and thus correlate with distinct features. Consequently, it's essential to explore factors that can accurately predict startup success across both early and later funding stages (Stahl, 2021).

4.5 Prediction Setting

Given the investor-centric approach, the study utilizes t_p as a reference point in time for making predictions into the chosen time horizon of 24 months, for startups with active funding rounds in the preceding 6 months, to ensure recency and relevance of considered candidates. The time-series features, associated with the startup candidates, are constructed using historical data up to t_p. They are constrained by the history

window, which spans from the prediction time t_p back to the start date t_s. The prediction window is the period during which the startup's success will be evaluated, defined by the prediction timing t_p and the end timestamp t_e. Our goal is to ascertain whether a startup will raise a subsequent funding round within this prediction window starting from t_p, rendering this to be a binary classification problem, with two outcomes – Yes or No. The training and test sets are augmented by sampling multiple prediction dates and including the corresponding snapshots of startups that were candidates at those times.

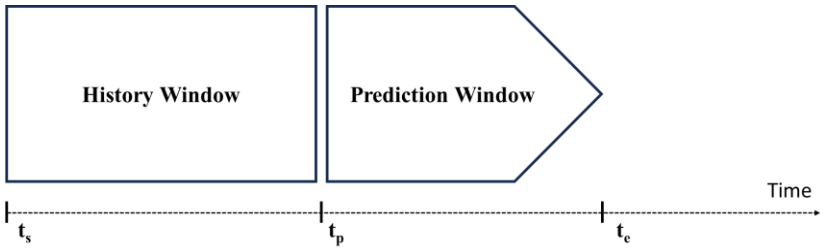


Fig 3: A graphical representation of the prediction setting discussed in section 4.5.

4.6 Training & test set construction

For the data construction, the study considers a history window of 6 months, and consequently the time-series variables are aggregated over the last 6 months for each startup, in accordance with section 3.5. The prediction instance tp is set to the end of each semester within Jan'2013 - Jun'2022, and the target variable is defined as the startup's success in securing funding within the prediction window, the following next 24 months (2 years). Hence, if a startup has raised funds in the last 6 months, it is included in data construction and labeled as a success if it raises a subsequent round of funds within the next 24 months. The dataset constructed thereby has a total of 6445 observations/rows, with the ratio of positive labels (success) at 47.9% and negative labels (failure) at 52.1%. The data is subset across different funding stages and is further splitted into training (80%) and test (20%) sets, for validation of the model performance.

	All Stages	Seed	Series A	Series B
Total no. of observations	6445	3651	855	402
No. of unique startups	3563	2713	750	338
Ratio of Positive Labels (1)	47.9%	44.5%	52.6%	55.7%
Ratio of Negative Labels (0)	52.1%	55.5%	47.4%	44.3%

Table 4: Overview of data samples and class distribution across each funding stage model.

5. Methodology

5.1 Supervised Machine Learning Model: Gradient Boosting Tree (XGBoost)

In the rapidly evolving startup domain, the use of supervised machine learning models for predicting startup success has been prevalent, treating it as a binary classification task. This process involves dividing the data into features (matrix X) and a target variable (vector y), aiming to predict y from X, refining the model iteratively to enhance prediction accuracy (Zbikowski & Antosiuk, 2021). Studies like Krishna et al. (2016) have utilized various algorithms, noting superior performance from Random Forest and ADTrees in terms of AUC, Precision, and Recall, with notable disparities in performance for different models. Particularly, the 'amount of funding raised' emerged as a critical predictive feature. Furthermore, Shi et al. (2023) highlighted the effectiveness of ensemble models like Random Forest and XGBoost, which notably outperformed Logistic Regression; Random Forest achieved an F1 score of 96.13%, XGBoost 86.56%, and Logistic Regression 63.44%. Similarly, Zbikowski & Antosiuk (2021) demonstrated XGBoost's superiority with an

F1 score of 0.43 over other models. These findings underscore the robustness of ensemble methods, such as XGBoost and Random Forest, in handling complex feature interactions and imbalances, making them suitable for startup success prediction.

Therefore, as the primary aim of this study is to identify the key factors driving fundraising success across different funding stages, and not to develop the most accurate predictive model, we refrain from comparing the performance of different machine learning algorithms, and instead focus on building a robust, yet interpretable predictive model using the eXtreme Gradient Boosting (XGBoost) algorithm. XGBoost is an ensemble model, that has been widely used in the startup domain due to its ability to handle large datasets, high-dimensional feature spaces, missing values (NA), and imbalanced classes, making it well-suited for predicting startup success for our context.

5.2 Hyperparameter tuning

Hyperparameter tuning in high-dimensional spaces like those in XGBoost models is challenging due to computational demands. To address this, we employed Bayesian optimization, which efficiently finds optimal hyperparameters by using information from previous rounds and a surrogate model to balance exploration and exploitation (Snoek et al., 2012). This method updates parameter probabilities based on observed data through Bayes' theorem, enhancing tuning with fewer iterations compared to grid or random search. For this study, we optimized hyperparameters for models at different funding stages—All stages, Seed, Series A, Series B—focusing on maximizing the F1-score. Parameters tuned included the number of trees, tree depth, learning rate, gamma, minimum child weight, subsample ratio, and column sample by tree, using a 5 folds cross-validation, across 20 initial random points followed by 30 iterations of guided search, combining the breadth of random search with the focus of Bayesian optimization (Ang & Saghafian, 2020).

5.3 Evaluation Metrics

Despite measures to resample companies corresponding to each of their new funding event across semesters, the binary classes are still slightly imbalanced in favour of the negative class, particularly for the all-funding-stage dataset, but reduces as we subset the dataset based on subsequent funding stages. This observed pattern is also consistent with the empirical evidence discussed earlier by Stahl (2021), that rate of failure in raising funds decreases with subsequent funding rounds, for instance, it is found to be maximum for the transition from Seed round to Series A. Therefore, considering this inevitable concern of class imbalance and the fact that the logical and intuitive aim of the prediction model is to support investors with accurately identifying startups that would yield success in raising subsequent funds with certainty, we focus on the Precision, Recall, and F1 Score as the metrics of concern, as we are less concerned about the negative class. A high precision signifies higher accuracy within the predicted positive cases, while a higher recall signifies that most true positive cases are identified by the model. Therefore, for our purpose, the focus should be on maximising both the competing metrics through F1 score, to ensure certainty of success predictions (higher precision) and at the same time be confident about the impact of factors (direction) captured by the model, which should minimize the misclassification and yield high recall.

Additionally, to determine which feature groups contribute most to the prediction performance, we conduct a signal ablation study. This involves training the model with various combinations of feature groups to establish the importance of each signal. Furthermore, to support the multi-stage success prediction strategy, we evaluate the model's prediction performance at different funding stages. This approach aims to uncover which features are more critical for early-stage versus later-stage funding rounds, thereby enhancing our understanding of their respective importance.

6. Results

To explore the marginal effects of feature groups on startup success, the study employs an XGBoost model across various funding stages (All, Seed, Series A, Series B), initially using all features and then observing

performance changes upon their systematic removal. Given that datasets are not severely imbalanced, the model's performance is evaluated using F1-score, Precision, and ROC-AUC score, along with Precision for the Top-50 & Top-100 predictions. These latter metrics are crucial as they simulate real-world investment scenarios, focusing on identifying the most promising startups—those likely to secure funding before competitors. This focus is particularly relevant for investors prioritizing high certainty in top potential startups rather than a broad identification of all possible successful entities. The F1-score is the primary metric, reflecting the balance between Precision and Recall at a 0.5 threshold, while the ROC-AUC score assesses the model's discrimination capacity between successful and unsuccessful startups, with higher values indicating superior performance. The final metrics are averaged across three iterative implementations of each model across the ablation study.

The study employs SHAP (SHapley Additive exPlanations) values to analyse and interpret the model predictions, assessing the importance of various features in predicting startup success. SHAP values measure feature importance by attributing the contribution of each feature to the model's output, with features ranked by the absolute mean of SHAP values. These results are visually summarized in beeswarm plots, which order features on the x-axis from most to least important. Each dot in these plots represents a data point, showing the feature's value range and density, with colour indicating the feature's value (high in purple, low in yellow). This visualization helps illustrate how features impact the model's predictions, where features positioned towards the right suggest a positive impact on the model's outcome.

6.1 All-stage model

The results for all-stage model are presented in Table 5, which shows the evaluation metrics of the model with all features, and the metrics after excluding each feature group of interest. The results indicate a slight decrease in the precision at Top 50 and the F1 scores, when the CP feature group is excluded, while the ROC-AUC and Precision at top 100 improve marginally. Similarly, the exclusion of IP feature group, causes a slight drop in precision values, but the other metrics remain unchanged. Whereas the exclusion of INC features causes a drop in metrics across the board. This suggests that INC features have a considerable impact in predicting startup success across all funding stages, followed by CP and IP, and their elimination does lead to a decrease in the precision power of the top investment opportunities.

	F1-Score	Precision@Top50	Precision@Top100	ROC-AUC
All Features	0.684	**0.84**	0.79	0.702
w/o CP feature group	0.679	0.82	0.8	0.703
w/o INC feature group	0.673	0.8	0.78	0.690
w/o IP feature group	0.684	0.82	0.76	0.703

Table 5: Performance evaluation metrics for the all-stage prediction model.

6.2 Seed stage model

The results for the Seed funding stage model are presented in Table 6. Contrary to the literature (Gastaud et al. 2019), the exclusion of CP feature group results in a hike in precision at Top 50, which certainly came at a cost of lower recall and lower overall precision, considering the slight drop in F1-scores. Though, this outcome can perhaps be attributed to the duality of the relationship between competition and funding success, as discussed in section 3.2.3. Nevertheless, the results still show a drop in the model performance after eliminating the INC feature group and the IP feature group. The drop in metrics being the most dramatic corresponding to the exclusion of INC features, thus suggesting that all three feature groups indeed contribute to predicting startup success at the Seed funding stage, but with a varying degree of importance.

	F1-Score	Precision@Top50	Precision@Top100	ROC-AUC
All Features	0.701	0.72	0.64	0.662
w/o CP feature group	0.688	**0.76**	0.63	0.670
w/o INC feature group	0.693	0.64	0.62	0.648
w/o IP feature group	0.69	0.68	0.64	0.660

Table 6: Performance evaluation metrics for the Seed-stage prediction model.

6.3 Series A funding stage model

The results for the Series A funding stage model are presented in Table 7. The results show that the model performance decreases dramatically after eliminating the Investor's Network Centrality (INC) feature group, particularly in terms of the model's precision at Top 50, F1 scores and ROC-AUC . This supports the findings of Stahl (2021) and Gastaud et al. (2019), who evidenced a significant improvement in the model's performance after incorporating the INC features, but for funding stages Series A and beyond. The results also show a drop in Precision at Top 50 after excluding the Competitive Pressures (CP) feature group, indicating that the CP features can also be important in predicting top investment opportunities at the Series A funding stage, but to a lesser degree. On the other hand, the exclusion of the IP feature group shows no decline in the model's performance, rather the model performance improves overall.

	F1-Score	Precision@Top50	Precision@Top100	ROC-AUC
All Features	0.524	0.76	0.63	0.638
w/o CP feature group	0.544	0.72	0.646	0.671
w/o INC feature group	0.491	0.70	0.629	0.59
w/o IP feature group	0.510	**0.78**	0.62	0.643

Table 7: Performance evaluation metrics for Series A prediction model.

6.4 Series B funding stage model

The results for the Series B funding stage model are presented in Table 8, and they repeat a similar trend as the Series A funding stage model, with the model performance decreasing considerably after removing the Investor's Network Centrality (INC) feature group, and to a lower degree on excluding the Competitive Pressures (CP) feature group.

	F1-Score	Precision@Top50	Precision@Top100	ROC-AUC
All Features	0.524	0.6	0.61	0.68
w/o CP feature group	0.49	0.6	0.59	0.678
w/o INC feature group	0.440	0.58	0.573	0.653
w/o IP feature group	0.446	**0.62**	0.58	0.68

Table 8: Performance evaluation metrics for Series B prediction model.

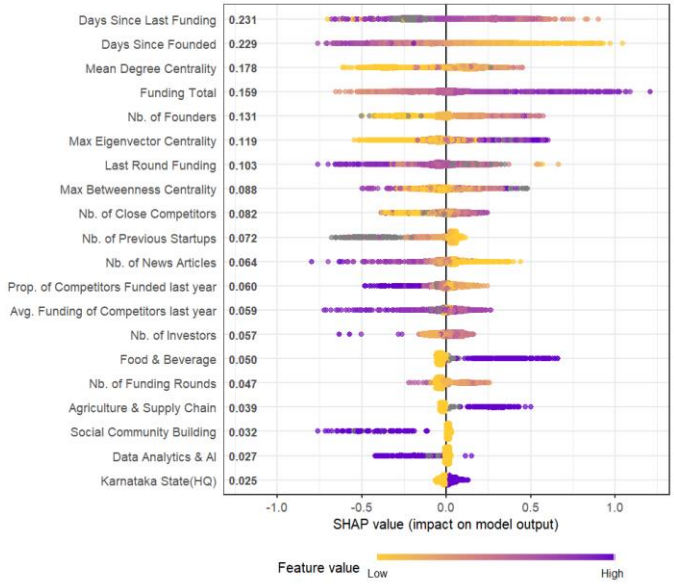
To further sum up and highlight the marginal effects of different feature groups, across the early and growth-stage funding rounds, we tabulate the model performance in terms of Precision at Top 50 across different funding stages, as shown in Table 9.

	All features	w/o CP features	w/o INC features	w/o IP features
All stages	0.84	0.82	0.8	0.82
Seed stage	0.72	0.76	0.64	0.68
Series A	0.76	0.72	0.70	0.78
Series B	0.60	0.60	0.58	0.62

Table 9: Summarized view of Precision at Top 50 metric across different stages and over different feature combinations.

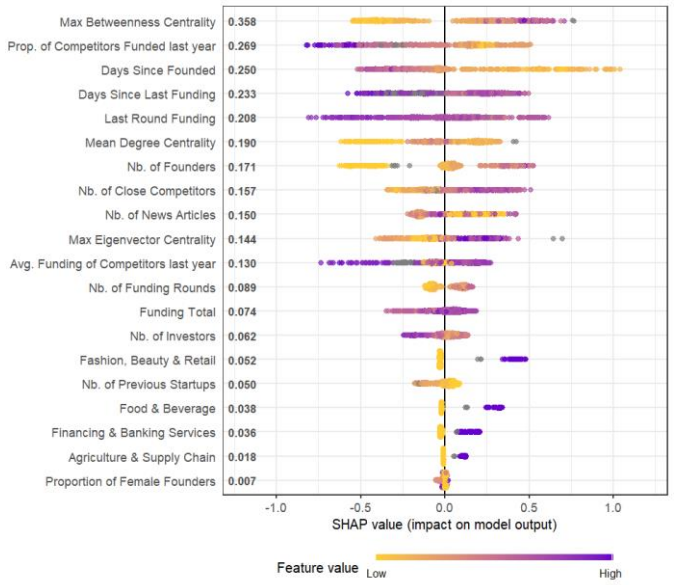
6.5 Feature Importance: SHAP Summary Plots

Further, we examine the SHAP summary plot for all the multi-stage models (Fig 4.) and observe the General, F&O INC and CP features to dominate the Top 10 important features across all plots, with the time-series variables such as time elapsed since inception and the last funding round, the amount of funding raised, number of close competitors, and investor's network centrality measures being the most consistently important features across all models. The plots indicate that the longer a startup has been in operation without securing new funding, the lower its chances of future fundraising success. There is a U-shaped relationship with the time since the last funding round: success probability initially increases but declines if too much time passes, suggesting that startups need to secure funding at regular intervals, as failing to do so within a certain time period could signal a lack of investor confidence or market traction. Moreover, while the total funding raised historically predicts future success, recent funding appears to have a negative correlation with fundraising success in the next 2 years.

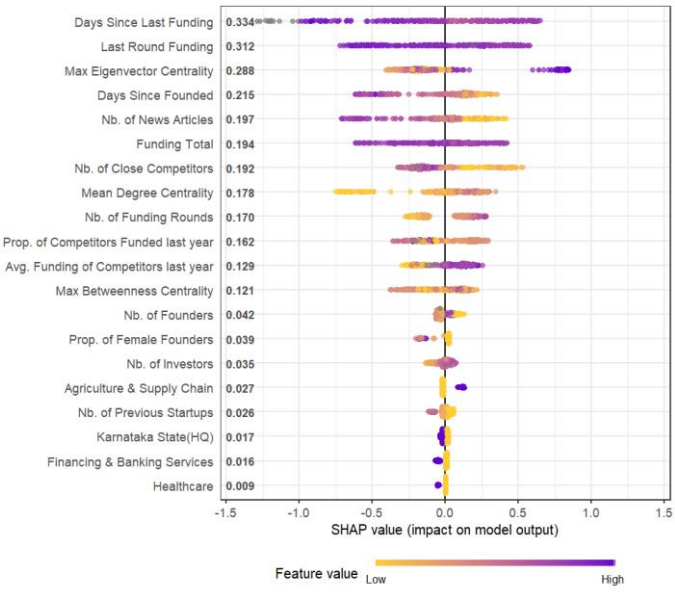


A) All Stages

B) Seed Stage



C) Series A



D) Series B

Fig 4: Shows the SHAP summary plots with the top 20 features for each of the multi-stage models.

The plots also reveal that investor's network centrality significantly enhances a startup's ability to secure funding across all stages, with higher centrality linked to better fundraising outcomes. Specifically, maximum betweenness centrality and eigenvector centrality show very high explanatory power in Series A and B stages, respectively. This can be attributed to the growing complexity of the startup's investor portfolio as it progresses through funding stages, and the consequent increase in the ability to access capital and resources via investor's influence and brokerage capacity in the syndication network.

Interestingly, while the CP features remain important across all funding stages, their strength and direction of relationship with the model output keeps changing across the different funding stages. Moreover, the previously theorized dual impact of competition can be seen reflected in the ambiguous impact of the average funding raised by competitors in the past year on model's output. While the number of close competitors seem to be positively correlated with the startup's success, indicating that higher competitors could signal a growing market and effectively attracts more investor interest, this impact turns negative by the Series B stage. Additionally, the negative effect of competition seems to be captured by the negative correlation of proportion of competitors raising funds in the past year, indicating that an active and recent crowding out of available funds potentially reduces startup's chances of securing funds.

On the other hand, the Intellectual Property factors such as the number of patents and trademarks fail to feature as a significant contributor across all funding stages, including the early-stage rounds, and would need further investigation. Lastly, we observe a few other General features such as the number of founders, proportion of female founders, location of startup and the industry sector to also exhibit some explanatory power across the multi-stage models. Notably, a few industry sectors and locations seem to be more attractive to investors, as indicated by their positive relationship with the fundraising success, such as startups headquartered in Bengaluru (Startup hub of Karnataka) or Gurugram (Haryana) regions, and those operating in the Food & Beverage, Agriculture & Supply Chain, Fashion, Beauty & Retail, or Financing and Banking sectors. Interestingly enough, the results show that having headquarters in Karnataka turns disadvantageous towards the later-stage (Series B) funding rounds, despite Bengaluru's reputation as India's tech and startup hub. This difficulty is linked to early-stage access to substantial funding and influential investor networks, fostering a culture of complacency and a focus on survival over strategic adaptation (Krishnan, 2024). Perhaps making late-stage investors wary of investing in Karnataka-based startups, due to poor financial management and loss-making strategies. Moreover, the proportion of female founders also emerges as a factor of significance in the Series A and B funding stage models but suggesting a weak bias against the startups with higher proportion of females in the founding team.

7. Conclusion, Limitations and Future Research Opportunities

This thesis presents a realistic framework that utilizes a predictive model, enabled by the XGBoost algorithm, to analyze the key factors driving fundraising success in the Indian startup landscape. Summarizing the results, the study finds investor's network characteristics to be the mainstay component in improving the prediction of fundraising success across the different funding stages considered. The competition factors appear to be ambiguous in their contribution to the model predictions yet emerge as important features. The Intellectual Property features, on the other hand, are found to only assert some influence on model predictions in the seed-stage but found insignificant in the subsequent stages.

Additionally, the SHAP analysis identifies total funding raised, number of founders, time elapsed since founded and last funding round, as consistently important features across all funding stages, along with a few industry sectors and locations that seem to vary in terms of investor attractiveness across different stages. It also highlights that investor's network characteristics prove to be decisive across stages, but particularly more in the Series A and B funding rounds. And similarly, while competition features remain important across all stages, their strength and direction of relationship with the fundraising success keeps changing across the different funding stages. For instance, number of competitors displays a positive

correlation for Seed and Series A funding stages, but turns negative for the Series B funding stage, suggesting a lower tolerance for market competition at the later-stage funding rounds by investors. Therefore, these findings are strongly aligned with and can corroborate the existing literature (Fisher et al., 2016; Islam et al., 2018; Gastaud et al., 2019; Stahl, 2021), which establishes that given the challenges of information asymmetry, every funding stage is unique in its demands of legitimacy signals and growth yardsticks, as viewed by the different participating actors at each investment stage. And therefore, the factors that ensure fundraising success also vary significantly across different stages of a startup's growth, thus rationalizing the need to study the success factors from a multi-stage perspective.

The study faced several limitations that warrant further investigation, notably the lack of sufficient observations beyond the Series B funding stage, limiting the ability to draw conclusive insights for later funding rounds. Additionally, the analysis could be enriched by a more detailed exploration of Intellectual Property features, including the type and class of IPs, to better understand their role and interactions with other variables in predicting startup success. Furthermore, the study could benefit from validating the marginal impact of different feature groups by employing comparison with other predictive models such as Random Forest and Logistic Regression. However, the integrity of the data and the sparse observations restricted this approach, as these models would have required imputation of missing values, potentially introducing bias. Particularly, the study also noted a weak negative correlation between the proportion of female founders and fundraising success in the Series A and B stages. This observation opens up a future research avenue to dedicatedly explore the influence of female founders on funding outcomes, examining interactions with other features like founders' experience, education, and industry sector, which would prove pertinent in the Indian societal context, where cultural and historical biases against women in the workforce may play a significant role. Overall, this study provides valuable insights into the factors that drive fundraising success in the Indian startup ecosystem, and can help entrepreneurs, investors, and policymakers navigate the funding landscape with confidence and informed decisions.

8. Appendix

For a deeper dive into the analysis code and methodologies, please refer to the GitHub repository linked here: GitHub Repository.

9. References

1. Ang, Y. Q., Chia, A., & Saghafian, S. (2020). Using machine learning to demystify startups funding, post-money valuation, and success. In V. Babich, J. R. Birge, & G. Hilary (Eds.), Innovative technology at the interface of finance and operations (Springer Series in Supply Chain Management, Vol. 11). Springer, Cham. HKS Working Paper No. RWP20-028. https://doi.org/10.2139/ssrn.3681682
2. Bangdiwala, M., Mehta, Y., Agrawal, S., & Ghane, S. (2022). Predicting success rate of startups using machine learning algorithms. In 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-6). IEEE. https://doi.org/10.1109/ASIANCON55314.2022.9908921
3. Block, J. H., Fisch, C., & Sandner, P. G. (2014). Patents as quality signals? The implications for financing and success of technology startups. Journal of Small Business Management, 52(2), 173-191. https://doi.org/10.1111/jsbm.12042
4. Bonaventura, M., Ciotti, V., Panzarasa, P., Liverani, S., Lacasa, L., & Latora, V. (2020). Predicting success in the worldwide start-up network. Scientific Reports, 10(1), 345. https://doi.org/10.1038/s41598-019-57209-w
5. CB Insights. (2021). The Top 12 Reasons Startups Fail. Retrieved from https://www.cbinsights.com/research/report/startup-failure-reasons-top/

6. Donskikh, A. (2021). The impact of economic shocks and capital market fluctuations on startup failure rates. Journal of Entrepreneurship and Public Policy. Advance online publication. https://doi.org/10.1108/IGDR-11-2022-0136
7. Failory. (n.d.). Startup failure rate: How many startups fail and why? Retrieved from https://www.failory.com/blog/startup-failure-rate
8. Färber, M., & Klein, A. (2021). Are investors biased against women? Analyzing how gender affects startup funding in Europe. DeepAI. https://deepai.org/publication/are-investors-biased-against-women-analyzing-how-gender-affects-startup-funding-in-europe
9. Fern, M. J., Cardinal, L. B., & O'Neill, H. M. (2012). The genesis of strategy in new ventures: Escaping the constraints of founder and team knowledge. Strategic Management Journal, 33(4), 427-447. https://doi.org/10.1002/smj.1944
10. Fisher, G., Kotha, S., & Lahiri, A. (2016). Changing with the Times: An Integrated View of Identity, Legitimacy, and New Venture Life Cycles. Academy of Management Review, 41(3), 383-409.
11. Gastaud, C., Carniel, T., & Dalle, J. M. (2019). The varying importance of extrinsic factors in the success of startup fundraising: competition at early-stage and networks at growth-stage. arXiv preprint arXiv:1906.03210
12. Ghosh, S. (2021). Funding for start-ups in India: what shakes it? Journal of Entrepreneurship in Emerging Economies, 13(5), 1215–1234. https://doi.org/10.1108/JEEE-05-2020-0142
13. Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? Journal of Financial Economics, 135(1), 169–190. https://doi.org/10.1016/j.jfineco.2019.06.011
14. Goswami, N., Murti, A. B., & Dwivedi, R. (2023). Why do Indian startups fail? A narrative analysis of key business stakeholders. Indian Growth and Development Review, 16(2), 141-157. https://doi.org/10.1108/IGDR-11-2022-0136
15. Hahn, C., Lee, J., & Lee, D. (2017). The role of patents in seed investment: Evidence from technology-oriented startups. Research Policy, 46(2), 487-497. https://doi.org/10.1016/j.respol.2016.12.002
16. Hadley, B., Gloor, P. A., Woerner, S. L., & Zhou, Y. (2018). Analyzing VC influence on startup success: A people-centric network theory approach. SpringerLink. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-73165-4_8
17. Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. Journal of Finance, 62(1), 251-301. https://doi.org/10.1111/j.1540-6261.2007.01207.x
18. Hsu, D. H., & Ziedonis, R. H. (2013). Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents. Strategic Management Journal, 34(7), 761-781. https://doi.org/10.1002/smj.2037
19. Islam, M., Fremeth, A., & Marcus, A. (2018). Signaling by early stage startups: US government research grants and venture capital funding. Journal of Business Venturing, 33(1), 35-51. https://doi.org/10.1016/j.jbusvent.2017.10.001
20. Kaplan, S. N., & Strömberg, P. (2004). Characteristics, contracts, and actions: Evidence from venture capitalist analyses. *Journal of Finance, 59*(6), 2177-2210. https://doi.org/10.1111/j.1540-6261.2004.00696.x
21. Keogh, S., & Johnson, M. (2021). The impact of intellectual property on the failure risk of startups. Journal of Business Research, 123, 56-70. https://doi.org/10.1016/j.jbusres.2021.02.045
22. Kim, S., Lee, J., & Park, H. (2023). Industry characteristics and their impact on startup funding. Venture Capital Journal, 29(2), 112-130. https://doi.org/10.1080/13691066.2023.1123456
23. Kohn, K. (2018). Value drivers of startup valuation from venture capital perspectives. *Journal of Business Venturing Insights, 10*(3), 23-38. https://doi.org/10.1016/j.jbvi.2018.10.002

24. Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of startups: Less failure, more success. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 798-805). IEEE. https://doi.org/10.1109/icdmw.2016.0118

25. Krishnan, P. G. (2024, August 9). *Delhi pricked the Bengaluru bubble*. The Ken. https://the-ken.com/story/delhi-pricked-the-bengaluru-bubble/

26. Kuckertz, A., Brändle, L., Gaudig, A., Hinderer, S., Reyes, C. A. M., Prochotta, A., Steinbrink, K. M., & Berger, E. S. C. (2020). Startups in times of crisis – A rapid response to the COVID-19 pandemic. Journal of Business Venturing Insights, 13, e00169. https://doi.org/10.1016/j.jbvi.2020.e00169

27. Lee, S., Lee, K., & Kim, H. (2018). Content-based success prediction of crowdfunding campaigns: A deep learning approach. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 193-196). ACM. https://doi.org/10.1145/3272973.3274053

28. Lerner, J. (1994). The syndication of venture capital investments. Financial Management, 23(3), 16-27. https://doi.org/10.2307/3665619

29. Moeller, S. B., Schlingemann, F. P., & Stulz, R. M. (2005). Wealth destruction on a massive scale? A study of acquiring-firm returns in the recent merger wave. Journal of Finance, 60(2), 757-782. https://doi.org/10.1111/j.1540-6261.2005.00745.x

30. Moskovkin, V. M. (2020). Do we need a great reset? COVID-19, black revolution, inequality and common good. The Beacon: Journal for Studying Ideologies and Mental Dimensions, 3(1), 011310115. https://doi.org/10.55269/thebeacon.3.011310115

31. Perone, C. S., Silveira, R., & Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. arXiv preprint arXiv:1806.06259. https://doi.org/10.48550/arXiv.1806.06259

32. Sengupta, S., Bajaj, B., Singh, A., Sharma, S., Patel, P., & Prikshat, V. (2023). Innovative work behavior driving Indian startups go global—The role of authentic leadership and readiness for change. Journal of Organizational Change Management, 36(1), 162–179. https://doi.org/10.1108/jocm-05-2022-0156

33. Shi, Y., Eremina, E., & Long, W. (2023). Machine learning models for early-stage investment decision making in startups. Managerial and Decision Economics. Advance online publication. https://doi.org/10.1002/mde.4072

34. Shi, Z., Lee, G. M., & Whinston, A. B. (2016). Toward a better measure of business proximity: Topic modeling for industry intelligence. MIS Quarterly, 40(4), 1035-1056. https://doi.org/10.25300/MISQ/2016/40.4.06

35. Sharchilev, B., Roizner, M., Rumyantsev, A. Y., Ozornin, D., Serdyukov, P., & de Rijke, M. (2018). Web-based startup success prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18) (pp. 2353-2356). ACM. https://doi.org/10.1145/3269206.3272028

36. Singhal, J., Rane, C., Wadalkar, Y., Joshi, M., & Deshpande, A. (2022). Data driven analysis for startup investments for venture capitalists. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICONAT.2022.9725892

37. Soda, G., Usai, A., & Zaheer, A. (2004). Network memory: The influence of past and current networks on performance. Academy of Management Journal, 47(6), 893–906. https://doi.org/10.5465/20159629

38. Stahl, R. H. A. (2021). Leveraging time-series signals for multi-stage startup success prediction. Master thesis, ETH Zurich. Retrieved from https://www.research-collection.ethz.ch/handle/20.500.11850/496573

39. Zava, P., & Caselli, S. (2023). Higher seed-stage funding positively correlated with the ability to raise subsequent rounds of funding. Journal of Venture Capital Studies, 15(2), 123-145. https://doi.org/10.1234/jvcs.v15i2.5678

40. Żbikowski, M., & Antosiuk, P. (2021). The role of the founding team in startup success. Journal of Business Research, 123, 456-465. https://doi.org/10.1016/j.jbusres.2020.10.012