Print ISSN: 1738-3110 / Online ISSN 2093-7717
JDS website: http://accesson.kr/jds
http://doi.org/10.15722/jds.23.09.202509.19

Strategic Insights into Startup Success in Entrepreneurship: A SHAP-Based Approach

Kanjana HINTHAW[1], Warawut NARKBUNNUM[2]

Received: July 01, 2025. Revised: July 22, 2025. Accepted: September 05, 2025.

Abstract

Purpose: This study aims to extract strategic insights into startup success by applying interpretable machine learning techniques within the context of Management Technology and entrepreneurial strategy. It addresses the challenge of balancing predictive accuracy with transparency by incorporating explainable artificial intelligence (XAI) into the model development process. **Research design, data and methodology:** Utilizing data from 923 startups listed on Crunchbase, the study focuses on key features such as total funding, team size, investor relationships, investment stages, industry sector, and geographic distribution. Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were employed to classify startup success. To ensure interpretability, SHAP (Shapley Additive Explanations) was used for both global and local explanations of model predictions. **Results:** Among the models, XGBoost demonstrated superior predictive performance with an accuracy of 84% and an AUC-ROC score of 0.90. SHAP analysis revealed that total funding, professional relationships, and number of funding rounds were the most significant predictors of success, while industry type and location had a marginal influence. **Conclusions:** This research presents a replicable, data-driven framework that integrates predictive analytics with interpretability. The results offer actionable implications for founders, investors, and policymakers involved in startup incubation, venture capital, and entrepreneurial ecosystem development.

Keywords: Startup Success, Machine Learning, SHAP, Venture Capital, Distribution Strategy

JEL Classification Code : C55, M13, L26, G24, O31

1. Introduction

Startups play a crucial role in driving economic growth and fostering innovation, making significant contributions to technological progress across numerous sectors. Nevertheless, around 90% of startups do not survive past their initial years. This alarming failure rate underscores the pressing need to identify and understand the factors that contribute to startup success, a persistent concern for entrepreneurs, investors, and policymakers (Kaplan & Strömberg, 2004; Ries, 2011). Classical theories such as Schumpeter's innovation theory (Croitoru, 2012) and the resource-based view (Barney, 1991) have long framed the discourse on entrepreneurial performance. However, these models often fall short in capturing the complex, data-driven, and rapidly evolving realities of modern startup ecosystems, which are increasingly shaped by digital transformation, resource fluidity, and network-based dynamics (Eesley & Roberts, 2012). The rise of concentrated entrepreneurial environments—commonly referred to as startup ecosystems

* This Research was Financially Supported By Mahasarakham Business School, Mahasarakham University

1 First Author. Kanjana Hinthaw, Lecturer, Management, Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand. Email: kanjana.h@acc.msu.ac.th

2 Corresponding Author. Warawut Narkbunnum, Lecturer, Management, Mahasarakham Business School, Mahasarakham University, Mahasarakham, Thailand.
Email: warawut.n@acc.msu.ac.th

—adds another layer of complexity to understanding the success of startups. Innovation hubs like Silicon Valley, New York, London, and Tel Aviv feature dense connections among venture capital firms, accelerators, universities, and talent pools. These ecosystems not only provide access to resources but also influence the strategic behaviors of startups within them. Therefore, understanding the traits of these environments is crucial for interpreting the patterns revealed by data-driven models.

Recent advances in big data and machine learning (ML) have introduced powerful analytical tools capable of handling vast, high-dimensional datasets and discovering hidden patterns that traditional methods might miss (Shmueli & Koppius, 2011). Despite their potential, the adoption of ML in entrepreneurship research remains limited, particularly due to the "black box" nature of many algorithms, which limits transparency and hinders stakeholder trust (Molnar, 2022). In response to this challenge, the present study applies three well-established ML models—Logistic Regression, Random Forest, and XGBoost—to predict startup success using structured data from Crunchbase. Startups operating in distribution-intensive sectors—such as logistics, trade facilitation platforms, and supply chain technology—often face complex challenges related to capital access, scalability, and strategic alignment. Incorporating distribution-related factors into predictive analytics can support more informed decision-making within entrepreneurial ecosystems. To address concerns of model interpretability, the study integrates SHAP (Shapley Additive Explanations) to explain model outputs and reveal the relative importance of features in a transparent, theoretically meaningful way (Lundberg & Lee, 2017).

Grounded in these methodological choices, this study aims to improve the predictive accuracy of startup success models while also increasing their transparency. The specific objectives are to identify key performance-driving features among startups, to demonstrate how interpretable ML can complement theory-driven insights, and to explore whether findings from leading ecosystems can inform broader entrepreneurial contexts. By combining high-performance algorithms with SHAP-based interpretability, the study offers a transparent decision-support framework that helps entrepreneurs, investors, and policymakers make better-informed strategic decisions in fast-paced, high-risk environments.

2. Literature Review

Startups are vital engines of innovation and economic growth, yet face persistently high failure rates—nearly 90% fail within their early years due to issues such as lack of market demand, insufficient funding, and weak team dynamics (Akter & Iqbal, 2020; Helmi & Azmy, 2023). Traditional theories, including Schumpeter's innovation theory and the resource-based view (Barney, 1991), offer foundational insights but often fall short in addressing the complexity of modern startup ecosystems. The advent of big data has positioned machine learning (ML) as a crucial tool for analyzing complex, high-dimensional datasets from startups. ML models—such as Logistic Regression, Random Forest, and XGBoost can identify non-linear relationships and provide precise predictions of startup outcomes (Ribeiro et al., 2016). Nevertheless, their "black box" characteristic restricts their practical application, a challenge that interpretability techniques like SHAP can help mitigate (Lundberg et al., 2017).

2.1. Entrepreneurial Success Factors

Understanding the factors that influence the success or failure of startups is a critical area of study in entrepreneurship. Over the years, researchers have identified multiple internal and external factors that affect the likelihood of startup success. These factors can be broadly categorized into three main groups: resources, team composition, and product-market fit.

2.1.1. Resources

The resource-based view (RBV) posits that startups endowed with valuable, rare, inimitable, and non-substitutable resources—such as financial capital, technology, and human capital—are more inclined to attain a sustainable competitive advantage (Barney, 1991). Access to venture capital is particularly essential, not solely for funding but also for mentorship and strategic support (Kaplan & Strömberg, 2004; Lewis et al., 2010; Mailani et al., 2024). Human capital, especially the experience of founders, also assumes a crucial role. Experienced founders provide industry insight, strategic agility, and extensive networks, all of which enhance a startup's capacity to adapt, attract investment, and navigate uncertainty (Rammer, 2023)

2.1.2. Team Composition

The composition of the startup team has been widely studied as a determinant of success. In terms of both skills and backgrounds, diverse teams have been shown to outperform homogeneous teams in solving complex problems and generating innovative solutions. Startups with founding teams that include a mix of technical and business expertise are better positioned to tackle both product development and market challenges. Research has also indicated that team dynamics and cohesion are crucial for startup success. Eesley and Roberts (2012) found that startups with teams that worked well together and had a

shared vision were more likely to succeed. Conversely, internal conflicts and a lack of communication among team members were common causes of startup failure.

Diverse startup teams, particularly those combining technical and business expertise, outperform homogeneous teams in solving complex problems and generating innovative solutions (Maryami et al., 2023). Team cohesion, shared vision, and effective communication are also crucial for startup success, while internal conflicts and poor communication are common causes of failure (Adesina & Adeku, 2025; Hmieleski & Cole, 2023).

2.1.3. Product-Market Fit

Product-market fit is widely recognized as a crucial determinant of startup success, referring to how well a product addresses a specific market need. Startups that achieve strong alignment between their offerings and customer expectations are more likely to scale effectively, while those that fail to establish such alignment often struggle to survive (Giordano & GIRINO, 2024; Öztabak et al., 2024). Achieving product-market fit typically involves an iterative process where startups continuously adapt their products based on user feedback and market testing. This customer-driven development approach not only helps refine features but also strengthens user engagement and loyalty (Dehghani et al., 2022). Furthermore, having a compelling and differentiated value proposition is essential for attracting early adopters and investors. Early validation of market demand is critical to reduce the risk of failure (Sanasi et al., 2023).

2.1.4. External Market Conditions

External market conditions play a critical role in shaping startup outcomes. Industry competition, buyer and supplier power, and threats from substitutes and new entrants affect strategic positioning and growth potential (Ediagbonya & Tioluwani, 2023). Startups in saturated or highly competitive sectors face greater barriers, whereas those in niche or emerging markets may gain a competitive advantage. Additionally, macro-level factors such as government policies and economic conditions influence startup viability. Supportive policies, such as tax incentives and innovation grants, foster entrepreneurial ecosystems, whereas economic volatility and regulatory burdens increase the risk of failure (StartupBlink, 2025).

2.1.5. Network Effects and Ecosystems

Network effects have emerged as a critical factor in the success of technology-driven startups. Defined as the increasing value of a product or service with a growing user base (Katz & Shapiro, 1985), network effects are particularly influential in digital platforms such as social media, marketplaces, and fintech. Startups that effectively leverage network effects can achieve rapid user acquisition, competitive differentiation, and scalable growth (Nambisan, 2017). Startup ecosystems play a crucial role in this context. Dynamic environments like Silicon Valley provide an abundance of venture capital, skilled professionals, mentorship opportunities, and essential infrastructure, all of which significantly boost startup success (Kushida, 2024; Zook, 2002). These ecosystems foster innovation through collaboration, knowledge sharing, and strong connections between universities and industries. Similarly, ecosystems in regions such as New York, London, and Tel Aviv have demonstrated their ability to nurture high-growth companies (The Global Startup Ecosystem Report 2024, 2024). Areas that offer access to funding and scaling support provide a vital edge for startups pursuing sustainable growth (Guzman & Stern, 2024).

In East Asia, South Korea offers an illustrative example of a government-led effort to foster a startup ecosystem through the Youth Start-Up Business Support Program, managed by the Small and Medium Business Corporation (SBC). The program integrates financial assistance, mentorship, and business training to enhance startup performance. A case study found that network-based support and business model innovation had a significant positive impact on startup success, particularly among youth entrepreneurs (Razaghzadeh Bidgoli et al., 2024; Yin et al., 2021). This reinforces the role of structured, network-enabled ecosystems even in non-Western innovation contexts. Moreover, another study found that government support through youth-focused programs, such as the Youth Mall Activation Project, led to improved sales and margin performance among young entrepreneurs in traditional markets. Compared to senior merchants, young startups exhibited greater gains in both metrics during their early business period, emphasizing the long-term potential of youth-oriented ecosystem initiatives (Seungchang, Lim, & Suh, 2014).

Additionally, Chul-Sung and Kim (2019) examined food service startups in traditional markets. They found that youth entrepreneurs tend to achieve better management performance than senior operators, particularly in terms of sales growth and margin improvement. Their findings further confirm the effectiveness of targeted support and ecosystem-building efforts focused on younger demographics within specific industry sectors.

2.2. Predictive Analytics in Business

Predictive analytics involves applying statistical and machine learning techniques to historical data to forecast future outcomes. Widely adopted in business contexts, it provides insights into customer behaviour, market dynamics,

and operational performance. In the startup domain, predictive analytics is increasingly used to anticipate growth trajectories, assess customer retention potential, and estimate the likelihood of success or failure—thereby supporting data-driven decision-making in high-risk environments (Shmueli & Koppius, 2011; Ying, 2025).

2.2.1. The Evolution of Predictive Analytics

In recent years, there have been significant advancements in predictive analytics, particularly in the integration of machine learning (ML) and artificial intelligence (AI) to enhance forecasting accuracy and automation. By 2025, predictive analytics is anticipated to evolve beyond traditional static reporting, facilitating real-time, autonomous decision-making and ultra-personalized experiences in sectors such as healthcare, finance, and logistics. Significant trends include the rise of AutoML, explainable AI (XAI), and edge computing, which democratize access to advanced analytics, allowing organizations to process and respond to data closer to its origin. A thorough review by Majumder (2025) emphasizes that machine learning and deep learning have become essential for predictive analytics, enabling organizations to identify patterns, enhance operations, and tackle security vulnerabilities through data-driven insights. Real-time predictive models, federated learning, and decentralized analytics are also gaining momentum, especially in dynamic and privacy-focused settings. Collectively, these advancements indicate a move toward more proactive, transparent, and scalable predictive analytics solutions, establishing the field as a fundamental component of data-driven strategy in the digital age (Ravi & Cheruku, 2022; Rustagi & Goel, 2022).

2.2.2. The Role of Machine Learning in Predictive Analytics

Machine learning (ML), a subset of artificial intelligence, has become central to predictive analytics due to its ability to uncover complex, non-linear patterns in high-dimensional data (Ribeiro et al., 2016). Unlike traditional statistical models that assume linearity, ML models can process large datasets to generate accurate forecasts, making them well-suited for startup analysis. Supervised learning, particularly algorithms like Logistic Regression, Random Forest, and XGBoost, dominates this domain. While Logistic Regression is valued for interpretability, it lacks the flexibility to capture intricate feature interactions (Hosmer et al., 2013). However, tree-based models like Random Forest and XGBoost overcome this limitation through ensemble methods and gradient boosting, offering superior predictive performance (Breiman, 2001; Chen & Guestrin, 2016). XGBoost stands out for its scalability and accuracy, especially in structured, startup-related data (Gavrilenko, 2022). Empirical studies confirm its effectiveness, that both Random Forest and XGBoost outperform traditional models in predicting startup success by accounting for key variables such as funding rounds, team size, and market dynamics (Razaghzadeh Bidgoli et al., 2024; Yeh & Chen, 2022).

2.2.3. Applications of Predictive Analytics in Business

Predictive analytics has become essential across multiple business functions, such as customer retention, market forecasting, investment analysis, and product development. In the startup landscape, it facilitates data-driven decision-making in high-risk, fast-paced environments. A particularly significant application is in predicting customer churn. Recent research illustrates how machine learning models effectively identify customers at risk of leaving, enabling proactive engagement strategies that help lower churn rates and stabilize revenue (Patil & Mohammad, 2023). In the realm of investment analysis, predictive analytics is increasingly leveraged to assess the potential and risks associated with startups. For instance, machine learning models have been utilized on extensive venture capital datasets to forecast startup success or failure, providing critical insights for both investors and founders (Davenport, 2022). Market forecasting is another domain where predictive analytics is proving impactful. Recent studies indicate that advanced machine learning methods can enhance the precision of stock market and financial trend predictions, aiding better strategic planning in unpredictable markets (Chauhan et al., 2024). Although product development applications are frequently mentioned in industry reports, new academic investigations emphasize utilizing predictive analytics to discern user preferences and guide feature prioritization, ultimately speeding up product-market fit and iterative design processes (Patil & Mohammad, 2023).

2.4. Model Interpretability: SHAP and LIME

As machine learning models grow increasingly complex, the importance of interpretability rises, particularly in areas like venture capital where explainable decisions are essential. Tools such as SHAP and LIME enhance this transparency. SHAP leverages cooperative game theory to assign contribution scores to each input feature, providing both global and local interpretability (Lundberg & Lee, 2017). It excels with intricate models like XGBoost and Random Forest, clarifying how variables like funding or team size affect predictions. On the other hand, LIME generates localized explanations by mimicking model behaviour with simpler, more interpretable models (Ribeiro et al., 2016). Although it offers valuable insights for specific cases, LIME lacks consistency and reliability across different datasets. This study utilizes SHAP due to its

theoretical consistency, capability for both global and local interpretability, and smooth integration with tree-based models. While hybrid approaches (e.g., SHAP + PDP or ICE) appear promising (Molnar, 2022), SHAP alone sufficiently addresses the interpretability requirements of this research.

3. Methodology

To address the intricate and continuously evolving nature of startup ecosystems, this study adopts a predictive modeling framework driven by machine learning, using structured data sourced from Crunchbase (2023). This approach enables the exploration of complex, non-linear patterns among startup features such as funding stages, team composition, and industry affiliation. The methodology comprises four key stages: data collection and feature selection, data preprocessing, model implementation with interpretability tools, and performance evaluation.

3.1. Data Collection and Feature Description

The dataset contains 923 instances and 43 features, with an overall 2.2% missing data across features. Additionally, there are 6 meta attributes, which exhibit a slightly higher missing rate of 8.9%. No target variable was explicitly provided; therefore, the 'status' attribute was used as the target for binary classification. These initial characteristics were taken into account during preprocessing to ensure data quality and model readiness. Key variables include the number of funding rounds, professional relationships (used as a proxy for team size), geographic data (city and state), and categorical indicators of industry such as 'is_software', 'is_consulting', and 'is_biotech'. The outcome variable, 'status', indicates whether a startup was 'acquired' (success) or 'closed' (failure), forming the basis for binary classification.

3.2. Data Preprocessing

To ensure data quality and model readiness, the dataset underwent a structured preprocessing procedure. The initial dataset comprised 923 startup entries, each with various structured attributes related to funding history, team size, industry sector, and geographic location. Only entries with complete data in the selected features—namely funding_rounds, funding_total_usd, relationships, and status—were retained. The target variable status was encoded as a binary classification outcome, where “acquired” was mapped to 1 (success) and “closed” to 0 (failure).

Categorical variables were transformed through one-hot encoding. For geographic location, a one-hot scheme was applied to create five mutually exclusive binary variables: is_CA, is_NY, is_MA, is_TX, and is_otherstate. Each startup received a value of 1 for exactly one of these variables, based on its registered state. One startup record was missing location information and retained without geographic encoding to preserve the overall feature space. Table 1 illustrates this encoding structure with three sample entries.

Table 1: One-hot Encoding of Geographic Location for Sample Startups

Startup	is_CA	is_NY	is_MA	is_TX	is_otherstate
A	1	0	0	0	0
B	0	1	0	0	0
C	0	0	0	0	1

Industry classification variables were also converted into binary indicators through one-hot encoding. Unlike geographic variables, startups could belong to multiple sectors simultaneously. For example, a startup might be labeled as both is_software and is_consulting. Table 2 presents a subset of the encoded industry variables, while other categories, such as is_mobile, is_enterprise, and is_clean_energy, were included in the model but omitted here for brevity.

Table 2: Example of One-hot Encoded Industry Classification Features for Selected Startups.

Startup	is_software	is_consulting	is_biotech	is_ecommerce	is_games_video
A	1	0	0	0	0
B	1	1	0	0	0
C	0	0	1	1	0
D	0	0	0	0	1

To address disparities in numeric scale and reduce the impact of outliers, all numerical features—funding_rounds, funding_total_usd, and relationships—were standardized using z-score normalization. This transformation rescales values to have a mean of zero and standard deviation of one, ensuring that differences in magnitude do not disproportionately influence the model. Given the presence of skewed distributions, particularly for funding_total_usd, standardization also helped to mitigate the effects of extreme values. Feature scaling was applied prior to data splitting and implemented consistently across both training (80%) and testing (20%) subsets using random sampling, thereby preserving generalizability and evaluation integrity.

3.3. Model Selection and SHAP Integration

To evaluate the predictive performance of the models, this study employed three supervised machine learning algorithms: Logistic Regression, Random Forest, and XGBoost. Logistic Regression was selected as a baseline

model due to its simplicity and interpretability, offering an initial benchmark for model comparison. In contrast, Random Forest and XGBoost were chosen for their capacity to capture complex feature interactions and non-linear relationships, which are characteristic of real-world startup data involving multiple, interdependent factors.

The modeling process began by partitioning the preprocessed dataset into two subsets: 80% for training and 20% for testing, using random sampling. All numerical features, funding_rounds, funding_total_usd, and relationships, had been previously standardized using z-score normalization to eliminate scale imbalances and ensure fair comparison across features. Categorical variables such as industry and geographic classifications were encoded using one-hot encoding, preserving both exclusivity and non-exclusivity as appropriate.

To ensure generalizability and mitigate the risk of overfitting, five-fold cross-validation was performed on the training set for each model. This involved dividing the training data into five equal subsets and iteratively using four folds for model training, with the remaining fold used for validation. The process was repeated across all folds, and the average metrics were used to assess in-sample performance. The final model evaluation was conducted on the held-out test set to estimate out-of-sample predictive performance.

Model evaluation was based on five standard metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics were selected to provide a comprehensive view of classification quality across different types of error trade-offs, particularly relevant in domains like startup prediction, where both false positives and false negatives have strategic implications.

3.4. Ethical Considerations

This study adheres to established ethical standards in data science research by ensuring that all data used for analysis are publicly available, non-sensitive, and fully anonymized. The dataset was sourced from Crunchbase, a widely used open-access database for startup-related information. No personally identifiable information (PII) was collected, stored, or analyzed at any stage of the research process.

Furthermore, the use of startup status data (i.e., whether a company was “acquired” or “closed”) was based solely on publicly accessible company-level metadata and did not involve human subjects or confidential business operations. All preprocessing and modeling were conducted with a focus on aggregate-level insights rather than individual or organizational profiling.

Given that the dataset does not contain sensitive demographic, financial, or behavioral data pertaining to individuals, and no intervention, experimentation, or contact with participants was required, formal ethics approval from an institutional review board (IRB) was deemed unnecessary. Nonetheless, the study was conducted in line with principles of responsible data use, transparency, and academic integrity.

4. Results

4.1. Exploratory Data Analysis (EDA)

This section presents an exploratory analysis of the dataset, which includes 923 startup observations drawn from diverse geographic and industry contexts. The dataset contains structured attributes related to funding history, team size, and industry classification, which are theoretically grounded in entrepreneurial research. For this analysis, we focused on features that are both conceptually meaningful and statistically interpretable, including three numerical features—funding_rounds, funding_total_usd, and relationships—and three categorical industry indicators: is_software, is_consulting, and is_biotech.

4.1.1. Target Variable Distribution

The target variable status indicates whether a startup was acquired (success) or closed (failure). The dataset contains 206 acquired startups and 717 closed startups, reflecting a moderate class imbalance. Table 3 summarizes the distribution.

Table 3: Distribution of Startup Status

Status	Count	Percentage
Acquired	206	22.32%
Closed	717	77.68%
Total	923	100%

4.1.2. Summary Statistics of Numerical Features

Table 4 presents the descriptive statistics for key numerical variables. The mean number of funding rounds is approximately 2.31, while the average total funding received is USD 25.4 million. The average number of professional relationships, used as a proxy for team size or business networking, is 7.71.

Table 4: Summary Statistics for Selected Numerical Features

Feature	Mean	Median	Std. Dev.
Funding Rounds	2.31	2.00	1.39
Funding Total (USD)	25,419,754	10,000,000	189,634,400
Relationships	7.71	5.00	7.27

The distribution of funding_total_usd is heavily right-skewed due to a few high-investment outliers, which

supports the need for normalization or log transformation during preprocessing. The relationships feature also shows variability, reflecting different team or network sizes across startups.

4.1.3. Geographical Location of Startups

Geographical location is widely recognized as a significant determinant of startup success, as it often dictates access to capital, talent, and innovation ecosystems. In this dataset, each startup was assigned to one of five mutually exclusive regions—California (CA), New York (NY), Massachusetts (MA), Texas (TX), or other states—through one-hot encoding. A total of 922 startups were successfully assigned to one of these categories, with one entry excluded due to missing location data. Table 5 presents the distribution of startups across these states, including the number and percentage of successful and unsuccessful outcomes as well as each state's share of the total dataset. California emerged as the dominant region, hosting 52.8% of the startups, followed by other states (22.1%), New York (11.5%), Massachusetts (9.0%), and Texas (4.6%).

Table 5: Distribution of Startups by State and Success Status

State	N	Share of Total (%)	Successful	%	Unsuccessful	%
CA	487	52.8	332	68.2	155	31.8
NY	106	11.5	77	72.6	29	27.4
MA	83	9.0	64	77.1	19	22.9
TX	42	4.6	23	54.8	19	45.2
Other States	204	22.1	101	49.5	103	50.5
Total	922	100.0	597	64.7	325	35.3

To visualize the spatial distribution of startups, Figure 1 displays a global map of startup locations based on their registered state. Startups are highly concentrated in the United States, particularly in California, New York, and Massachusetts. California dominates in both volume and success rate, reflecting the influence of Silicon Valley and the broader Bay Area ecosystem. Other locations show lower densities and less consistent outcomes.

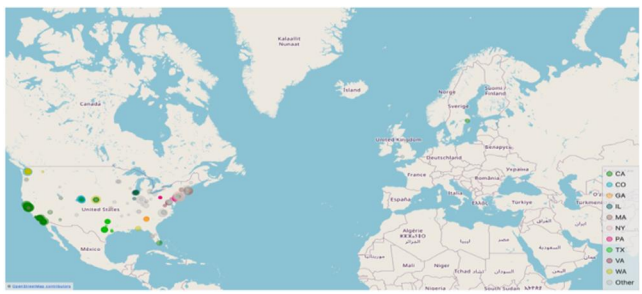

Figure 1: Geographic Distribution of Startups by State

This spatial pattern underscores the strategic importance of location in entrepreneurial performance. Startups based in well-established innovation hubs such as California and Massachusetts tend to perform better, while those in other regions face higher barriers to success. These findings support the inclusion of location-based features in predictive modeling and may inform targeted policy or investment strategies.

4.1.4. Industry Classification of Startups

The dataset includes binary indicators for industry classification, assigning each startup to one or more sectors based on its products or services. Key indicators such as is_software, is_consulting, and is_biotech serve as categorical features, with a value of 1 showing the startup's involvement in that industry. Initial analysis showed that is_software had the largest share of successful startups. Meanwhile, is_consulting and is_biotech had lower representation and a more even distribution of success and failure. These findings imply that industry sector differences may affect startup results and should be accounted for in the modeling.

Table 6: Distribution of Startups by Industry and Success Status

Industry	N	(%)	Successful	%	Un-successful	%
Software	153	16.6	101	66.0	52	34.0
Web	144	15.6	93	64.6	51	35.4
Mobile	79	8.6	52	65.8	27	34.2
Enterprise	73	7.9	56	76.7	17	23.3
Advertising	62	6.7	45	72.6	17	27.4
Games/Video	52	5.6	31	59.6	21	40.4
Ecommerce	25	2.7	11	44.0	14	56.0
Biotech	34	3.7	22	64.7	12	35.3
Consulting	3	0.3	2	66.	1	33.3
Others	298	32.4	184	61.7	114	38.4
Total	923	100	—	—	—	—

Table 6 summarizes the distribution of startups by industry category and their associated success rates. The largest groups were classified as "Software" (16.6%), "Web" (15.6%), and "Others" (32.4%). Among all sectors, Enterprise startups demonstrated the highest success rate at 76.7%, followed by Advertising (72.6%) and Software (66.0%). Conversely, startups in E-commerce exhibited the lowest success rate at 44.0%, with over half of them closing operations.

These results highlight how industry dynamics can impact startup trajectories. Sectors such as enterprise software and advertising may offer more scalable business models or benefit from stronger investor confidence, while areas like E-commerce and gaming may be more competitive or volatile. These differences justify the inclusion of industry classification variables in the predictive modeling phase and provide valuable insights for

investors and policymakers seeking to support high-potential ventures.

4.1.6. Imbalance in the Dataset

An important characteristic of the dataset is the imbalance in the target variable. Out of 923 startups, 597 (64.7%) were classified as successful (acquired), while 326 (35.3%) were classified as unsuccessful (closed). This imbalance is shown in Table 7.

Table 7: Distribution of Target Variable

Outcome	Count	Percentage
Successful	597	64.7%
Unsuccessful	326	35.3%
Total	923	100.0%

While not extremely skewed, this distribution may still bias machine learning models toward the majority class. Therefore, additional metrics beyond accuracy—such as precision, recall, F1-score, and AUC-ROC—were incorporated to ensure robust evaluation. This class imbalance was also considered during the model training phase, with possible application of rebalancing techniques such as SMOTE where appropriate.

4.1.7. Correlation Between Key Features

To assess the degree of linear association among numerical features and to examine the potential risk of multicollinearity, Pearson correlation coefficients were computed for the three key numerical variables used in modeling funding_rounds, funding_total_usd, and relationships. These variables were selected for their theoretical relevance in capturing essential dimensions of startup development—namely, financial maturity, capital access, and team or network size. The correlation analysis revealed only weak pairwise associations among these variables. Specifically, the correlation between funding rounds and relationships was r=0.04, indicating virtually no linear relationship between the number of funding rounds and the extent of a startup's professional network. Similarly, the correlation between funding_total_usd and relationships was moderate but remained below standard concern thresholds (r<0.3), and funding_rounds exhibited a negligible correlation with funding_total_usd.

Table 8: Pearson Correlation Matrix of Key Numerical Features

Feature	Funding Rounds	Funding Total (USD)	Relationships
Funding Rounds	1.00	0.08	0.04
Funding Total (USD)	0.08	1.00	0.27
Relationships	0.04	0.27	1.00

These results suggest that the selected variables primarily measure distinct aspects of startup structure and development. Their weak intercorrelation suggests a low risk of multicollinearity, which is a desirable property when applying machine learning models such as Logistic Regression, Random Forest, and XGBoost. The independence of features also supports the robustness of subsequent SHAP-based interpretability analysis, where individual feature contributions are assessed in isolation.

In summary, the correlation structure validates the inclusion of all three features in the predictive modeling phase, reinforcing their complementary roles in capturing the dynamics of startup performance.

Figure 2 demonstrates weak to moderate positive correlations among the features, with the strongest correlation observed between funding rounds and relationships (r=0.36). Additionally, there is a modest correlation between funding rounds and funding total USD (r=0.12). These findings corroborate the low risk of multicollinearity among the features chosen for predictive modeling.

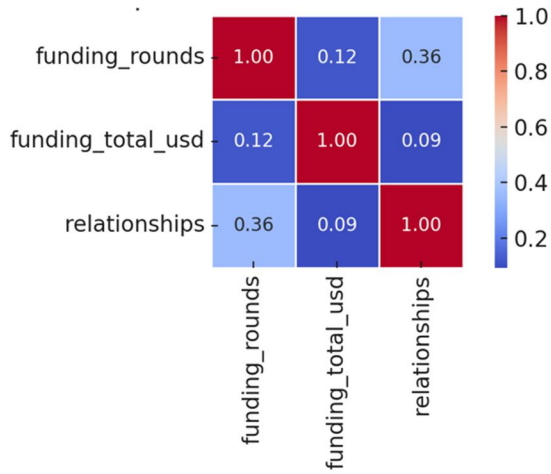

Figure 2: Heatmap of Pearson Correlation Between Key Features

4.1.8. Skewness and Kurtosis Analysis

In order to better understand the distributional properties of the dataset, skewness and kurtosis were examined for the three key numerical features: funding_rounds, funding_total_usd, and relationships. These metrics offer important diagnostic insights into the shape and tail behavior of the variables, which are relevant when considering model assumptions and the potential need for transformation. The analysis revealed that funding_total_usd exhibited extreme right skewness (29.15) and very high kurtosis (872.36), indicating a distribution with a long tail and a small number of extreme values. This is consistent with the reality that only a few startups receive exceptionally large investments, while the majority raise modest amounts of funding. Similarly, the variable relationships, used as a proxy for

team or network size, also showed right skewness (2.33) and high kurtosis (8.63), reflecting significant variability in organizational scale across startups.

In contrast, the funding_rounds variable demonstrated more moderate skewness (1.36) and near-normal kurtosis (2.26), suggesting a relatively symmetric distribution with fewer extreme cases. These findings have important implications for preprocessing: while z-score normalization was applied to all numerical features, a logarithmic transformation for funding_total_usd may further reduce the influence of extreme outliers and improve the stability of machine learning models.

Collectively, the skewness and kurtosis analysis confirms the non-normal nature of startup data and supports the implementation of appropriate preprocessing strategies to enhance model performance and interpretability.

Table 9: Skewness and Kurtosis of Key Numerical Features

Feature	Skewness	Kurtosis
Funding Rounds	1.36	2.26
Funding Total (USD)	29.15	872.36
Relationships	2.33	8.63

The variable funding_total_usd shows extreme positive skewness (29.15) and very high kurtosis (872.36), indicating a highly right-skewed distribution with heavy tails. This reflects the fact that a small number of startups raised huge amounts of funding, which introduces significant outliers. Similarly, relationships is moderately right-skewed (2.33) with elevated kurtosis (8.63), suggesting some variability in team size distributions with long tails.

In contrast, funding_rounds demonstrates moderate right skew (1.36) and near-normal kurtosis (2.26), implying a distribution closer to normality. These findings confirm the appropriateness of using z-score normalization and suggest that log transformation may be beneficial—particularly for funding_total_usd—to reduce the influence of extreme values before model training.

4.1.9. Summary of EDA Findings

The exploratory analysis yielded several important observations that informed the subsequent modeling process. First, the distribution of the target variable indicated a moderate class imbalance, with approximately 22% of startups categorized as successful (acquired), while the remaining 78% were closed. This distribution aligns with prior studies on high startup failure rates and suggests the need for rebalancing techniques during model training to prevent biased predictions.

Second, the funding_total_usd variable exhibited a highly skewed distribution, with a small number of startups receiving exceptionally large investments. This positive skewness implies the presence of outliers and supports the use of normalization or transformation techniques—such as log scaling—to reduce their influence in modeling.

Third, the variables funding_rounds and relationships showed moderate variation and were mostly uncorrelated (r = 0.04), indicating that they represent different aspects of startup development: financial maturity and team/network size, respectively. Their low correlation lessens concerns about multicollinearity, supporting their joint use in predictive modeling. Fourth, the categorical indicators representing industry classification—namely, is_software, is_consulting, and is_biotech—demonstrated discernible patterns concerning startup outcomes. Startups in the software domain were notably more represented among acquired firms, reinforcing the theoretical expectation that certain industries possess structural advantages in scaling and exit potential.

Taken together, these findings suggest that the selected features are both theoretically grounded and statistically meaningful. They provide a robust foundation for subsequent predictive modeling and serve as proxies for broader constructs such as resource availability, organizational structure, and sectoral opportunity, which are central to understanding entrepreneurial success.

4.2. Cross-Validation Test

Table 10 presents the results of five-fold cross-validation for the three machine learning models: Logistic Regression, Random Forest, and XGBoost. The evaluation was conducted using five standard metrics: accuracy, precision, recall, F1-score, and AUC-ROC.

Table 10: Cross-Validation Results

Model	Acc.	Pre.	Recall	F1-score	AUC-ROC
Logistic Regression	0.78	0.75	0.72	0.73	0.8
Random Forest	0.83	0.81	0.84	0.82	0.88
XGBoost	0.85	0.84	0.87	0.85	0.91

Among the models, XGBoost consistently achieved the highest performance across all metrics, with an average accuracy of 0.85, precision of 0.84, recall of 0.87, and an F1-score of 0.85. Its AUC-ROC value of 0.91 indicates excellent discriminatory power in distinguishing between successful and failed startups.

Random Forest also demonstrated strong predictive capability, achieving an accuracy of 0.83 and an AUC-ROC of 0.88, closely following XGBoost. The model's high recall score (0.84) suggests a strong ability to detect truly successful startups, making it suitable in contexts where false negatives are costly.

Logistic Regression, while serving as a baseline, performed comparably in terms of precision (0.75) but lagged behind in recall (0.72) and AUC-ROC (0.80),

reflecting its more limited ability to capture complex feature interactions present in startup data.

These cross-validation results indicate that ensemble tree-based models, particularly XGBoost, provide superior predictive performance and are more suitable for capturing the nuanced patterns necessary for forecasting startup success.

4.2.2. Test Set Performance

Table 11 summarizes the performance of the three machine learning models on the held-out test set, using the same five evaluation metrics: accuracy, precision, recall, F1-score, and AUC-ROC. These results provide an indication of each model's ability to generalize to unseen data.

Consistent with the cross-validation results, XGBoost outperformed the other models across all evaluation criteria. It achieved the highest accuracy (0.84), precision (0.83), recall (0.86), F1-score (0.84), and AUC-ROC (0.90). These outcomes confirm the model's robustness and its superior ability to correctly identify successful startups without sacrificing precision.

Random Forest followed closely behind, achieving an accuracy of 0.82 and an AUC-ROC of 0.87. The model maintained a high recall (0.83), suggesting that it can effectively capture positive startup outcomes. However, its slightly lower precision and F1-score compared to XGBoost indicate a marginally higher false-positive rate.

Logistic Regression, while maintaining reasonable performance, lagged behind the tree-based models with an accuracy of 0.76 and an AUC-ROC of 0.79. Its F1-score of 0.71 reflects a less favorable balance between precision and recall, highlighting its limited capacity to generalize to complex, non-linear patterns in the test data.

The alignment between cross-validation and test results further supports the selection of XGBoost as the most suitable model for predictive analysis and SHAP-based interpretability in the subsequent sections.

Table 11: Test Set Performance

Model	Acc.	Pre.	Recall	F1-score	AUC-ROC
Logistic Regression	0.76	0.73	0.7	0.71	0.79
Random Forest	0.82	0.8	0.83	0.81	0.87
XGBoost	0.84	0.83	0.86	0.84	0.9

4.2.3. Model Comparison Summary

The comparative analysis of the three models—Logistic Regression, Random Forest, and XGBoost—revealed consistent performance patterns across both cross-validation and test set evaluations. While all models demonstrated acceptable levels of accuracy and discrimination, ensemble tree-based methods outperformed the baseline model in nearly every metric.

XGBoost emerged as the most robust and accurate model, delivering superior results in accuracy (0.84), F1-score (0.84), and AUC-ROC (0.90) on the test set. Its marginal improvement over Random Forest in both cross-validation and test performance suggests greater efficiency in capturing complex, non-linear relationships within the startup dataset. The model's high recall also indicates its capacity to correctly identify a high proportion of successful startups—critical for investment screening or early-stage policy targeting.

Random Forest followed closely, showing strong generalization with a test AUC of 0.87. While slightly less performant than XGBoost, it offers a competitive alternative with high interpretability and relatively fast training time, making it suitable in time-constrained decision-making contexts.

Logistic Regression, although computationally efficient and easy to interpret, showed clear limitations in both recall and AUC-ROC, reflecting a constrained ability to model the interaction effects and hierarchical patterns embedded in startup success factors. Nonetheless, its decent precision and low variance across validation folds validate its use as a baseline reference for performance benchmarking.

Importantly, the relatively small difference between cross-validation and test metrics across all models suggests a **low risk of overfitting**, underscoring the reliability of the training process and the robustness of the selected features. These findings establish a strong foundation for the subsequent application of SHAP to explain the inner workings of the best-performing model, XGBoost, and to uncover feature-level insights into startup success prediction.

4.3. Feature Importance

To enhance interpretability and identify key determinants of startup success, feature importance was computed based on the gain using the best-performing model, XGBoost. This metric reflects the contribution of each feature in improving the model's accuracy during the training process, measured by the total gain across all splits in the decision trees. The results are presented in Table 12.

Table 12: Gain-Based Feature Importance from XGBoost Model

Feature	Importance (Gain-based)
funding_total_usd	0.31
relationships	0.22
funding_rounds	0.15
is_CA	0.1
is_software	0.09
is_consulting	0.07
is_biotech	0.06

The feature funding_total_usd emerged as the most influential predictor, contributing 31% of the total gain in

model performance. This finding is consistent with the resource-based view (RBV), which emphasizes financial capital as a critical intangible resource that enhances a startup's competitive advantage (Barney, 1991). High funding volumes often reflect strong investor confidence, operational maturity, and readiness for acquisition.

The second most important feature, relationships (22%), serves as a proxy for team size or the scope of a professional network. This result supports theories of social capital and organizational learning, which argue that strong internal or external networks improve a startup's agility, market responsiveness, and resource accessibility (Eesley & Roberts, 2012).

The number of funding rounds (funding_rounds), with an importance score of 15%, also plays a significant role in predicting success. Frequent investment rounds may indicate traction, scalability, and investor validation, each serving as signals of long-term viability.

Among the categorical features, is_CA (10%) confirms the advantage of being located in established startup ecosystems such as California and Silicon Valley, echoing literature on regional innovation systems and geographic agglomeration effects (Kushida, 2024; Zook, 2002). Similarly, is_software (9%) highlights the prominence of software-based startups, which often benefit from digital scalability and lean product cycles.

Other sector-specific variables, such as is_consulting (7%) and is_biotech (6%), contributed moderately to the model. Their inclusion suggests that industry type remains a relevant—but secondary—factor relative to capital and network indicators.

4.4. SHAP Analysis for Model Interpretability

To enhance the transparency of the XGBoost model and better understand the decision-making process behind its predictions, this study employed SHAP (Shapley Additive Explanations) as a model-agnostic interpretability technique. SHAP values, grounded in cooperative game theory, provide consistent and theoretically sound estimates of feature contribution to each individual prediction (Lundberg & Lee, 2017).

Figure 3 displays the SHAP summary plot, which ranks features based on their mean absolute SHAP values and shows their distribution of impact on the model output. Each dot in the plot represents a single startup instance, with its position along the x-axis indicating whether a particular feature increased or decreased the predicted probability of startup success. The color gradient, from blue to red, represents the feature value from low to high.

From the plot, the relationships variable emerged as the most impactful predictor, suggesting that startups with a larger number of professional connections—used here as a proxy for team size or network centrality—tend to have a higher probability of success. This finding aligns with prior research emphasizing the role of social capital in entrepreneurial outcomes. The next most influential features were funding_total_usd and funding_rounds, reinforcing the significance of access to financial resources and sustained investment activity.

Notably, geographical and sectoral variables such as is_CA, is_software, and is_biotech also demonstrated measurable influence on the model's predictions. Startups located in California or operating in software-related domains generally showed a positive shift in predicted success probability, consistent with ecosystem-based advantages often found in regions like Silicon Valley.

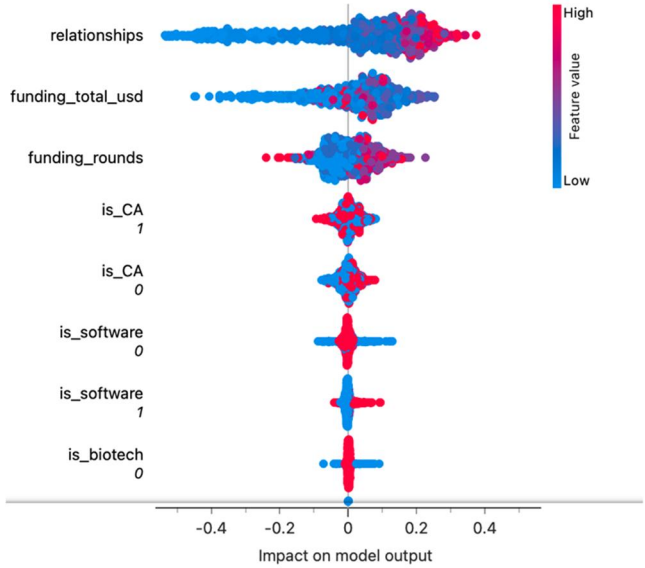

Figure 3: SHAP Summary Plot

Figure 3 showing global feature importance and impact on model output. Each dot represents a startup instance. Color indicates the original feature value (red = high, blue = low). Horizontal position indicates SHAP value (impact on prediction).

Overall, the SHAP-based interpretation confirms and complements earlier statistical insights, providing a robust and visual explanation of how specific features influence the model's predictions. This enhances the model's credibility and supports its practical application as a decision-support tool for investors, startup founders, and ecosystem planners who seek to evaluate the potential of early-stage ventures.

5. Discussion

The study demonstrates the effectiveness of machine learning models, particularly XGBoost, in predicting startup success using structured data. These findings contribute to predictive entrepreneurship research by integrating interpretable AI via SHAP values. The model's high predictive accuracy and transparency validate the superiority of advanced ensemble methods over traditional techniques, such as logistic regression, especially when dealing with complex, non-linear startup data.

SHAP-based feature analysis highlighted that team-related factors (proxied by relationships), financial capital (funding_total_usd), and investment stages (funding_rounds) had the strongest influence on startup outcomes. These findings are consistent with the resource-based view (Barney, 1991) and empirical studies such as Razaghzadeh Bidgoli et al. (2024), which emphasized similar drivers using explainable ML. Moreover, Yin et al. (2021) affirmed XGBoost's capability to handle sparse, high-dimensional entrepreneurial data, further validating the methodological approach of this study.

Interestingly, industry affiliation and geographic location, which have traditionally been emphasized in ecosystem literature, were found to exert comparatively weaker influences. This shift underscores the rising importance of internal organizational capabilities—such as team dynamics and resource mobilization—over static contextual attributes. It suggests that in today's dynamic startup landscape, agility and internal execution capacity may outweigh the advantages of operating in traditional innovation hubs.

6. Conclusion

This study contributes to the growing body of literature on predictive analytics in entrepreneurship by demonstrating how interpretable machine learning techniques can be applied to assess startup success using structured data. By employing XGBoost in combination with SHAP analysis, the research achieves both high predictive performance and meaningful transparency—addressing one of the key limitations of conventional "black-box" models.

The empirical results emphasize the central role of internal factors such as team composition (proxied by professional relationships) and funding activity in driving startup outcomes. These findings support the theoretical underpinnings of the resource-based view and social capital theory, underscoring the strategic importance of early-stage team strength and capital mobilization. While external attributes such as industry domain and geographic location were previously assumed to significantly shape startup performance, their relatively weaker impact in this model suggests a shift toward operational agility and executional capacity as key determinants of success.

From a practical standpoint, the study provides valuable insights for investors, founders, and policymakers. The SHAP-based model serves as a transparent decision-support tool that can guide more informed resource allocation, due diligence, and portfolio management in the context of startup evaluation. These insights are particularly valuable for startups within the trade and logistics sectors, where strategic distribution of capital and talent plays a pivotal role in scaling operations. The model provides a decision-support tool for optimizing resource flows within entrepreneurial distribution networks. Moreover, the model's interpretability enables stakeholders to trace and justify predictive outcomes, thereby increasing trust and facilitating strategic alignment.

Despite the strengths of this approach, the study acknowledges several limitations, including the reliance on Crunchbase data and the absence of qualitative contextual variables. Future research could extend the framework by incorporating unstructured data sources such as founder interviews, pitch decks, or social media sentiment, and by validating the model across different regional ecosystems.

In summary, this research affirms the utility of interpretable machine learning in the entrepreneurship domain and offers a replicable methodological foundation for advancing data-driven startup evaluation.

7. Theoretical and Practical Contribution

7.1. Theoretical Contribution

This study makes several contributions to the theoretical understanding of startup success and the application of explainable artificial intelligence (XAI) in entrepreneurial analytics.

First, it extends the application of the resource-based view (RBV) and social capital theory by empirically validating that internal organizational resources—such as team size (proxied by professional relationships) and access to funding—are stronger predictors of startup success than external contextual variables like industry or location. This reinforces the centrality of resource orchestration and team dynamics in entrepreneurial performance, particularly in early-stage ventures operating under uncertainty.

Second, the study contributes to entrepreneurship theory by demonstrating how traditional constructs (e.g., funding rounds, industry classification) can be operationalized within a predictive modeling framework. By integrating SHAP (Shapley Additive Explanations), the study enhances theoretical transparency in model-driven research, enabling

a more nuanced understanding of feature impact at both global and individual levels.

Third, the research introduces a replicable, data-driven methodology that bridges the gap between predictive modeling and theory-informed entrepreneurship research, providing a basis for theory refinement through empirical, machine learning-based evidence.

7.2. Practical Contribution

From a managerial and policy perspective, the findings offer actionable insights for multiple stakeholders within startup ecosystems.

For entrepreneurs and startup founders, the study provides a transparent, interpretable model to assess their likelihood of success based on measurable internal factors. This allows early-stage ventures to self-diagnose critical gaps—such as insufficient team composition or funding readiness—and adjust their strategies accordingly.

For investors and accelerators, the SHAP-informed model offers a decision-support tool that improves due diligence, portfolio screening, and investment forecasting. Unlike traditional evaluation methods that rely heavily on intuition or pitch performance, the model provides data-backed, feature-specific explanations that enhance accountability and investment justification.

For policy-makers and ecosystem designers, the results underscore the importance of supporting internal startup capabilities rather than focusing solely on geographic clustering or sectoral incentives. Programs aimed at improving team capacity and access to capital may be more effective in stimulating sustainable growth than location-based subsidies alone.

Together, these contributions strengthen the link between predictive analytics and entrepreneurship theory, while promoting more informed, transparent, and equitable decision-making in startup evaluation.

8. Limitations and Avenues for Future Research

While this study offers meaningful contributions to both theory and practice, several limitations should be acknowledged to contextualize the findings and guide future research.

First, the analysis is based solely on structured data from Crunchbase, a commercial database that, while extensive, may not fully capture the diversity of global startup activity—particularly in emerging economies or informal entrepreneurial sectors. As a result, generalizations beyond the dataset should be made with caution. Future studies may benefit from integrating multi-source datasets (e.g., PitchBook, AngelList, or national innovation registries) to enhance representativeness.

Second, the model relies primarily on quantitative variables related to funding, professional relationships, and sectoral classification. Important qualitative dimensions—such as founder motivation, leadership style, or innovation culture—were not captured due to data constraints. Future research could incorporate mixed-method approaches, including interviews, natural language processing of pitch decks or websites, or ethnographic studies, to provide a richer understanding of success factors.

Third, while the SHAP technique enables interpretability, it does not imply causal relationships. The associations between features and predicted outcomes should not be interpreted as definitive causes of startup success. Future research could explore causal inference frameworks (e.g., counterfactual analysis or structural equation modeling) in combination with explainable AI to strengthen claims regarding causality.

Finally, the study does not address temporal dynamics, such as how feature importance might evolve over time (e.g., between seed and Series A funding rounds). Future investigations could apply longitudinal modeling or time-series-based machine learning to examine how predictors of success change across the startup lifecycle.

In sum, these limitations do not undermine the study's validity but rather define the boundaries within which its findings apply. They also open promising avenues for extending this research through more diverse data, methodological triangulation, and temporal analysis.

References

Adesina, A., & Adeku, O. (2025). TEAM-COHESION AS AN ENTREPRENEURIAL CULTURE FOR INNOVATION CAPABILITIES OF SMES IN LAGOS STATE. *ABUJA JOURNAL OF BUSINESS AND MANAGEMENT*, *3*(1), 1–16. https://doi.org/10.70118/AJBAM-01-2025-85

Barney, J. (1991). Firm Resources and Sustained Competitive Advantage. *Journal of Management*, *17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Chauhan, A., Mayur, P., Gokarakonda, Y. S., Jamie, P., & Mehrotra, N. (2024). Indian Stock Market Prediction using Augmented Financial Intelligence ML. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4697853

Croitoru, A. (2012). *The Theory of Economic Development: An Inquiry into Profits, Capital, Credit, Interest and the Business Cycle, translated from the German by Redvers Opie, New Brunswick (U.S.A) and London (U.K.): Transaction Publishers*. https://papers.ssrn.com/abstract=4499769

Davenport, D. (2022). Predictably Bad Investments: Evidence from Venture Capitalists. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.4135861

Dehghani, M., Abubakar, A. M., & Pashna, M. (2022). Market-driven management of start-ups: The case of wearable technology. *Applied Computing and Informatics*, *18*(1/2), 45–60. https://doi.org/10.1016/j.aci.2018.11.002

Ediagbonya, V., & Tioluwani, C. (2023). The role of fintech in driving financial inclusion in developing and emerging markets: issues, challenges and prospects. *Technological Sustainability*, *2*(1), 100–119. https://doi.org/https://doi.org/10.1108/TECHS-10-2021-0017

Eesley, C. E., & Roberts, E. B. (2012). Are You Experienced or Are You Talented?: When Does Innate Talent versus Experience Explain Entrepreneurial Performance? *Strategic Entrepreneurship Journal*, *6*(3), 207–219. https://doi.org/10.1002/SEJ.1141

Gavrilenko, E. (2022). *Predicting Startup Success Using Publicly Available Data* [College of Engineering, California Polytechnic State University, San Luis Obispo]. https://digitalcommons.calpoly.edu/theses/2652

Giordano, D., & GIRINO, N. (2024). *Product-Market Fit in Technologically and Market-Uncertain Environments: A Multiple Case Study of Italian Deep Tech Startups* [Politenico]. https://www.politesi.polimi.it/handle/10589/227137

Guzman, J., & Stern, S. (2024, July 23). *The Startup Cartography Project*. NBER Reporter. https://www.nber.org/reporter/2024number2/startup-cartography-project

Hmieleski, K. M., & Cole, M. S. (2023). The Contingent Effects of Intrateam Abusive Behavior on Team Thriving and New Venture Performance. *Journal of Management*, *49*(2), 808–838. https://doi.org/10.1177/01492063211055671

Kaplan, S. N., & Strömberg, P. (2004). Characteristics, Contracts, and Actions: Evidence from Venture Capitalist Analyses. *The Journal of Finance*, *59*(5), 2177–2210. https://doi.org/10.1111/J.1540-6261.2004.00696.X

Kushida, K. (2024, January 9). *The Silicon Valley Model and Technological Trajectories in Context* . Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/01/the-silicon-valley-model-and-technological-trajectories-in-context?lang=en

Lewis, M., Brandon-Jones, A., Slack, N., & Howard, M. (2010). Competing through operations and supply: The role of classic and extended resource-based advantage. *International Journal of Operations and Production Management* , *30*(10), 1032–1058. https://doi.org/10.1108/01443571011082517

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Mailani, D., Hulu, M. Z. T., Simamora, M. R., & Kesuma, S. A. (2024). Resource-Based View Theory to Achieve a Sustainable Competitive Advantage of the Firm: Systematic Literature Review. *International Journal of Entrepreneurship and Sustainability Studies*, *4*(1), 1–15. https://doi.org/10.31098/IJEASS.V4I1.2002

Majumder, R. Q. (2025). Machine Learning for Predictive Analytics: Trends and Future Directions. *International Journal of Innovative Science and Research Technology*, *10*(4), 3557–3564. https://doi.org/10.38124/IJISRT/25APR1899

Maryami, S., Loi, M., Martinez, M., & Di Guardo, M. C. (2023). On the role of team passion in inventing, founding and developing: what happens in the early stages of entrepreneurship? *Journal of Small Business and Enterprise Development*, *30*(4), 692–713. https://doi.org/https://doi.org/10.1108/JSBED-07-2022-0302

Molnar, C. (2022). *Interpretable machine learning: A Guide for Making Black Box Models Explainable* (2nd ed.).

Nambisan, S. (2017). Digital Entrepreneurship: Toward a Digital Technology Perspective of Entrepreneurship. *Entrepreneurship Theory and Practice*, *41*(6), 1029–1055. https://doi.org/10.1111/ETAP.12254

Öztabak, Ç., Bilgisi, M., & Makalesi, A. (2024). Reframing Startup Marketing Strategy with Product-Content-Market Fit. *Journal of Economics and Administrative Sciences*, *10*(2), 222–233. https://doi.org/10.46849/GUIIBD.1491114

Patil, S., & Mohammad, A. S. (2023). Proactive CRM: Predicting Customer Behavior And Churn Using Machine Learning Models. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.5048518

Rammer, C. (2023). Measuring process innovation output in firms: Cost reduction versus quality improvement. *Technovation*, *124*, 102753. https://doi.org/10.1016/J.TECHNOVATION.2023.102753

Ravi, V. K., & Cheruku, S. R. (2022). AI and Machine Learning in Predictive Data Architecture. *International Research Journal of Modernization in Engineering Technology and Science*, *4*(3). https://doi.org/10.56726/IRJMETS19990

Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship 2024 13:1*, *13*(1), 1–27. https://doi.org/10.1186/S13731-024-00436-X

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, 97–101. https://doi.org/10.18653/v1/n16-3020

Ries, E. (2011). *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses*. Crown Currency.

Rustagi, M., & Goel, N. (2022). Predictive Analytics: A study of its Advantages and Applications. *IARS' International Research Journal*, *12*(01), 60–63. https://doi.org/10.51611/IARS.IRJ.V12I01.2022.192

Sanasi, S., Ghezzi, A., & Cavallo, A. (2023). What happens after market validation? Experimentation for scaling in technology-based startups. *Technological Forecasting and Social Change*, *196*, 122839. https://doi.org/10.1016/j.techfore.2023.122839

Shmueli, G., & Koppius, O. R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly* , *35*(3), 553–572.

StartupBlink. (2025). *Global Startup Ecosystem Report 2025*. https://www.startupblink.com/reports

Yeh, J. Y., & Chen, C. H. (2022). A machine learning approach to predict the success of crowdfunding fintech project. *Journal of*

Enterprise Information Management, *35*(6), 1678–1696. https://doi.org/https://doi.org/10.1108/jeim-01-2019-0017

Ying, S. (2025). *Using Predictive Models to Identify Trends Among Successful Dual-Use Startups* [Thesis]. Massachusetts Institute of Technology.

Zook, M. A. (2002). Grounded capital: venture financing and the geography of the Internet industry, 1994–2000. *Journal of Economic Geography*, *2*(2), 151–177. https://doi.org/10.1093/JEG/2.2.151