HIGH SCHOOL EDITION
Journal of Student Research

The Influence of Various Factors in a Startup's Success: A Comprehensive Study

Akshaj Nadimpalli[1] and Aliya Babul[#]

[1]South Windsor High School, USA
[#]Advisor

ABSTRACT

This study explores the many factors that contribute to the success of startups, examining a wide range of elements that can influence their outcomes. By analyzing over 4,000 cases, the research uses advanced statistical and machine learning techniques like Decision Trees, Random Forests, and Gradient Boosting to assess how well different startup traits can predict success. These traits include the industry, location, how long the company has been around, the number of previous ventures the founders have launched, and their educational backgrounds. The results reveal important predictors of success, shedding light on how these factors shape outcomes. With highly accurate predictive models, the study provides valuable insights for those involved in the startup world, offering a solid foundation for making informed decisions. Beyond contributing to the academic conversation on startup success, this research gives entrepreneurs and stakeholders practical, data-driven strategies to navigate the competitive startup environment.

Introduction

The entrepreneurial landscape is a fast-paced and ever-changing space where many factors come together to shape whether startups succeed or fail. Understanding what drives these outcomes is essential for entrepreneurs, investors, and policymakers as they navigate the challenges of building and sustaining new businesses. This study takes a deep dive into how different factors influence startup success, using a data-driven approach to uncover the complexities of the startup ecosystem. By analyzing a rich dataset and applying advanced analytical methods, the research looks closely at how industry sectors, location, company age, founder experience, and education impact outcomes. Through the use of cutting-edge machine learning techniques, the study not only identifies key predictors of success but also measures their effects, providing a detailed look at what really drives startup performance. Positioned at the crossroads of entrepreneurship research and predictive analytics, this study seeks to illuminate the factors behind startup success and offer practical insights to help create a thriving startup environment.

Literature Review

Synergy in Product, Process, and Innovation Management

A study by Habiburrahman et al. (2022) provides an in-depth look at the startup scene, focusing on 100 participants, including 41 incubators and 59 startups from Banyuwangi, Jember, Madiun, Malang, and Surabaya. The spread of these incubators and startups across different cities highlights the diversity of the startup ecosystem in the region. Banyuwangi and Malang are home to 24% of the incubators, while Banyuwangi leads with 27% of the startups, followed by Malang and Jember, each hosting 20%. In terms of business turnover, Madiun and Surabaya stand out, with revenues ranging from below IDR 300 million to IDR 4.5 billion. The study uncovers some key differences in what incubators and startups prioritize for success. Incubators tend to focus on product development, innovation, and processes, while startups put more emphasis on collaboration, communication, and innovation. These differences in focus reflect the varied approaches and strategies that contribute to startup success, depending on the regional context. (Habiburrahman et al., 2022).

Strategy and Innovation During Covid-19:

In a study by Damayanti, Yuswanto, and Hariri (2022), 219 individuals, mostly startup owners and workers, were surveyed to explore what drives startup success in the post-COVID-19 world. After filtering the data, 211 respondents were included in the final analysis. The researchers used Structural Equation Modeling (SEM) to examine the connections between key factors like strategy, innovation, experience, and economic growth, with 'Startup Success' acting as a central variable. The findings revealed that while strategic planning is essential for startup success and has a strong impact on economic growth, innovation and experience did not have a significant direct effect on success. This study underscores the importance of having a solid strategic plan to navigate the challenges of the post-pandemic business environment.

Government Policy and Financial Capital

Tan Le Trinh's 2019 study examines the key factors that influence the performance of small and medium-sized enterprises (SMEs) in Danang City. It highlights the crucial roles that government policies, financial resources, cultural and social influences, and human capital play in driving the success of these businesses. Using structural equation modeling with partial least squares (PLS-SEM), the research analyzes how these elements impact startup outcomes. The results suggest that having supportive legal frameworks and access to credit is especially important for startups during their early stages. This study offers valuable insights into how both external factors, like government support, and internal factors, like human capital, contribute to the success of SMEs in the region.

New Product Development Using Fuzzy Logic

AlHazza et al. (2019) explore the key success factors in new product development for startups using a Fuzzy logic approach. The study identifies several critical elements for success, including the uniqueness of the product idea, a well-defined business model, and an effective marketing strategy. It also highlights the challenges startups face, particularly the unpredictability and high risks involved in developing new products. The authors argue that traditional methods may not be ideal in such uncertain environments, where qualitative judgments and "fuzzy" factors come into play. By applying Fuzzy logic, the research provides valuable insights into how startups can navigate these complexities, manage their operations more effectively, and take advantage of technological opportunities.

Software Development Startup Dynamics

Shanbhag and Pardede's 2019 study examines the intricate process of product development in software startups, emphasizing the importance of using system dynamics as a tool to understand this process. The research highlights how time, capital, and product differentiation work together as critical factors in the success of software development startups competing in a fast-paced market. By employing system dynamics modeling, the study offers a three-dimensional view of these success factors, allowing for a deeper understanding of the complex interactions and challenges that software startups face. This approach provides valuable insights for strategic decision-making, helping startups navigate the highly competitive tech industry.

In their 2022 article, Shanbhag and Pardede expand on their earlier work by continuing to explore the dynamics of product development in software startups. Using causal loop constructs from system dynamics, the study visualizes the interactions between key factors that influence startup success. The research again identifies time, capital, and product differentiation as critical, three-dimensional success factors. Emphasizing the fierce competition faced by software startups, the study provides a valuable framework for understanding and navigating this challenging landscape. It sets the stage for more comprehensive future research in this field, offering a holistic view of the factors driving success in software development.

Entrepreneurial Ecosystems

Mai and Nguyen's 2022 study explores how entrepreneurial ecosystems shape entrepreneurs' perceptions and the success of their startups. Using survey data from 200 founders and CEOs of SMEs and startups in Tay Ninh City, Vietnam, the researchers applied the Partial Least Squares (PLS) method to analyze the data. The findings reveal that five out of six ecosystem factors significantly influence both entrepreneurs' perceptions and their startups' success, with the entrepreneurs' perceptions also having a positive effect on their success. This research highlights the crucial role of a supportive entrepreneurial ecosystem in fostering successful entrepreneurship and driving business growth at the national level.

Techno-Entrepreneurship Innovation

A 2022 study on techno-entrepreneurship in developing economies identifies seven Critical Success Factors (CSFs) that play a pivotal role in the success of technology-based entrepreneurial ventures. Based on a survey of 250 entrepreneurs who use technology platforms for their businesses, the research highlights how these CSFs enable entrepreneurs to either launch new ventures or transform existing businesses into virtual platforms. The findings show a strong positive relationship between these CSFs and the success of ventures, emphasizing their importance in driving growth and sustainability for startups in the tech sector.

EdTech Startups in Adult Education

Chavkin's 2020 study examines the factors that influence scalability for Russian EdTech startups, specifically in the field of adult education. The research identifies five key groups of factors tied to crucial elements of a business model: market, product, customer relations, distribution, team, and investments. The study challenges the idea that simply copying a successful competitor's business model or relying on a subscription-based model guarantees success. Instead, it finds that a global market orientation, leveraging key market trends, and focusing on high-quality distribution and promotion are far more critical to scaling and succeeding in the EdTech sector.

Effect Of Founders’ Education on Startup Success

Based on the data compiled in the study "Understanding the impact of entrepreneurial education on startup success" from SpringerLink, several key findings can be highlighted regarding the success factors in startup funding. The educational background of founders plays a crucial role, where higher levels of education generally correlate with increased success in securing funding. Interestingly, founders with education across multiple disciplines show more success in funding, but this trend reverses if the education is too broad. Startups with multiple founders are approximately 27% more likely to surpass early-stage hurdles.

In terms of specific educational backgrounds, 54.6% of founders with technical degrees have secured funding compared to 50.5% of those with business education. However, founders possessing both technical and business degrees have a higher success rate (37.2%) than those with only business degrees (30%). Previous experience, such as exiting a prior startup, also increases the likelihood of securing investments by 23%. The study further reveals that the rate of securing funding declines with lower educational degrees: 56.1% for doctorates, 47.0% for postgraduates, and 43.4% for undergraduates.

Significantly, the research identifies various factors that statistically influence startup investment success. These include gender, age, number of co-founders, experience, self-sustainability, co-founders from the same university, and cofounders' technical education - all significant at a 1% level. Cofounders with both business and technical education are significant at a 5% level, while startups funded by external organizations are significant at the 10% level. Interestingly, the study notes that doctoral level business education and undergraduate business education are significant at the 5% and 10% levels, respectively.

Furthermore, the research indicates that technical, business, and general degrees affect the probability of a startup being self-sustained and successfully funded. Co-founders with technical, business, or general degrees have a 9%, 4%, and 5% lower chance of being self-sustained, respectively, but a 7%, 3%, and 4% higher probability of being funded. Co-founders from the same university are less likely to be self-sustained but more likely to receive external funding or exit the startup. Additionally, previous startup experience increases the chances of receiving venture capitalist funds by 16% while decreasing the probability of being self-sustained. This comprehensive analysis sheds light on the multifaceted factors that influence the trajectory of startup funding and sustainability.

Methodology

The study began with the aggregation of a dataset designed to capture a diverse range of variables influencing startup success. The dataset included categorical and continuous variables, consisting of the following:

1. Details.Headquarters Regions: Different regions where companies are headquartered.
2. Sector: Various sectors or industries companies belong to.
3. Overview: General information about companies, including number of founded organizations and gender distributions.
4. Regions: Geographic regions related to the companies' operations or origins.
5. Education: Levels of education and whether they attended a prestigious university.
6. Major: Fields of study or majors of individuals associated with the companies.
7. Target: A binary target variable for analysis or prediction.
8. Details.Years Since Founded: Information about how long the companies have been founded.

To analyze the influence of these factors, the study employed a range of Python libraries specifically designed for data manipulation and analysis. Pandas and NumPy were used as the core tools for efficient data handling and numerical operations, respectively. For data visualization, Matplotlib and Seaborn were utilized to create detailed graphical representations, making it easier to identify and interpret the dataset's underlying patterns. When it came to predictive modeling, the study relied on Scikit-learn, a widely respected library known for its comprehensive machine learning tools. This allowed for a thorough exploration of the predictive dynamics within the startup ecosystem, providing deeper insights into the factors driving success.

Initial Data Examination

The investigation starts with a detailed examination of the dataset, using Pandas to load the data and inspect its structure and dimensions. This initial step is essential for gaining a comprehensive understanding of the dataset's scale, including the number of entries and features, as well as the types of variables it contains. By exploring this foundational layer, the study establishes a clear picture of the data's composition, allowing for a more nuanced and precise analytical process to follow. This step ensures that any inconsistencies or patterns within the data are identified early, laying the groundwork for deeper analysis.

Heatmap Creation Using Cramér's V

To explore the relationships between categorical variables, the study employs Cramér's V statistic, a specialized measure for gauging the strength of association between two nominal variables. The analytical process involves constructing a matrix of Cramér's V values, capturing the relationships between variable pairs, such as sector affiliations and headquarters locations. This matrix serves as the foundation for a heatmap visualization, carefully designed to offer a clear and quantitative depiction of these inter-variable associations. By distilling complex categorical relationships into an intuitive and visually engaging format, this technique adds a valuable layer of quantitative insight to the analysis, making it easier to identify patterns and correlations within the dataset.

Bar Graphs for Success Rates

To highlight differences in startup success rates across various categories, such as geographical location, the study uses bar graphs for their clear and straightforward visual comparison. Bar graphs make it easy to directly compare success rates across different groups, allowing patterns to emerge quickly. This part of the analysis focuses on drawing practical insights from these comparisons, making the differences in success rates obvious and easy to understand.

Scatter Plots and Regression Analysis

The study delves into continuous variables by using scatter plots combined with regression analysis to reveal trends and correlations. This approach blends the clarity of visual representation with the precision of statistical modeling, helping to explore relationships between variables like "Years Since Founded" and "Number of Founded Organizations." By adding regression lines to the scatter plots, the study provides a solid method for spotting trends, making it easier to identify patterns in the data with a statistical basis.

Box Plots for Distribution Analysis

To closely examine the distribution of variables, especially metrics like the years since founding among successful startups, the study uses box plots. This method is ideal for highlighting the spread of data, identifying outliers, and showing key insights into central tendencies and variability. The use of box plots is purposeful, aimed at breaking down the distribution patterns of important variables related to startup success. This approach helps provide a clearer understanding of the factors that shape startup growth and outcomes.

Predictive Modeling and Evaluation

Decision Tree Classifier

The study uses the Decision Tree classifier for its straightforward and easy-to-interpret structure. This model, much like a flowchart, maps out the decision-making process step by step, based on different attributes, ultimately leading to a classification result. The key strength of the Decision Tree is its non-parametric nature, meaning it doesn't rely on any assumptions about the data. By breaking down complex datasets into simple decision rules, this model offers an intuitive yet powerful way to analyze and understand the factors that predict startup success.

Random Forest Classifier

Using the Random Forest Classifier highlights the study's focus on ensemble learning for improved predictive accuracy and feature analysis. This model combines the results of multiple decision trees to make stronger, more reliable predictions, effectively handling the complexity and size of the dataset. One of the key strengths of Random Forest is its ability to reveal the importance of individual features, offering a clear, data-driven understanding of which variables are most influential in predicting startup success.

Gradient Boosting Classifier

The use of the Gradient Boosting Classifier showcases an advanced approach to predictive modeling, where each step builds on the previous one to improve accuracy. By optimizing loss functions and correcting errors in a step-by-step manner, this model represents some of the most sophisticated machine learning techniques in the study. Known for its precision and flexibility, the Gradient Boosting Classifier enhances the study's ability to make accurate predictions while also helping to prevent overfitting, ensuring that the model performs well across various data scenarios.

Logistic Regression

The study used machine learning models, including Logistic Regression, Random Forest, and Gradient Boosting, to predict startup success. Data preprocessing involved standardizing features and splitting the data into training and testing sets. Model performance was evaluated using accuracy, precision, recall, and F1-scores. Feature importance analysis identified key predictors, and a probability prediction function was developed for practical application.

Findings

Description Of Startup Landscape

Sector

The dataset's analysis on the distribution of startups across various industries reveals a pronounced preference for startups within certain sectors, with the category labeled as "Other" standing out due to its high concentration. This category, being the most populous, serves as an indicator of the diverse range of industries that startups are venturing into. However, the dominance of the "Other" category also introduces a layer of complexity to the analysis, as it encompasses a wide variety of sectors not explicitly defined within the traditional industry classifications. This aggregation under "Other" suggests that while it represents the largest group, the potential for more nuanced analysis within is somewhat obscured, given its broad and undefined nature. In descending order of startup counts, the sectors line up with "Technology" and "Health Care" following the "Other" category. These sectors, along with "Financial Services," demonstrate a significant presence in the startup ecosystem, indicating their attractiveness and potential for growth and innovation. The "Technology" and "Health Care" sectors, each with over 1000 startups, highlight the critical role these industries play in driving entrepreneurial endeavors. "Financial Services" trails closely, suggesting its importance in the startup landscape, albeit with slightly fewer startups compared to "Technology" and "Health Care." On the lower end of the spectrum, sectors such as "Industrials," "Energy," and "Real Estate" are characterized by their lesser startup counts with fewer than 500. This lower activity level points to a more modest but still significant engagement of startups within these industries, indicating diverse opportunities for innovation and growth across the board. The notable disparity in startup distribution, especially with the "Other" category's predominance, underscores the diversity within the startup ecosystem.

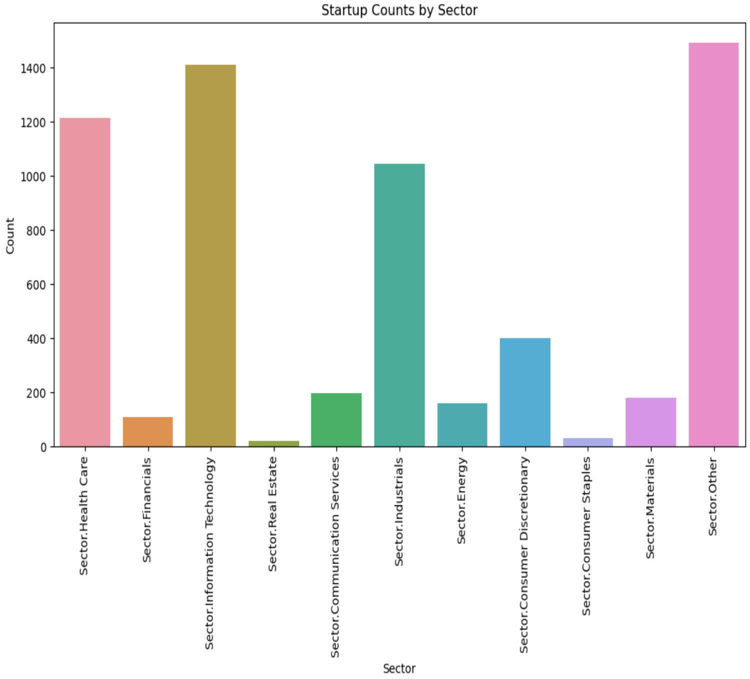


Figure 1. Shows the startup counts based on the different sectors.

Headquarters Location

The distribution of startup counts by headquarters location offers a visual comparison of startup distribution across various geographic regions. Although exact numerical values are not provided, visual estimations suggest that the "Greater New York" area or the "Great Lakes" region is the most common headquarters location for startups, potentially serving as the mode of the dataset. The range of startup counts, inferred from the highest to the lowest counts, is roughly estimated to be between 300-350 startups, with the "Greater New York" or "Great Lakes" exhibiting the highest concentration and the "New York Area" the lowest. An estimated ordering of locations by startup count presents "Great Lakes" at the forefront, closely followed by "Greater New York," with "Other" and "New England" in the middle range, and the "Midwestern US" and "New York Area" trailing. This ordering reflects a moderate variability in startup presence across locations, with "Great Lakes" hosting around 400 startups and the "New York Area" the least, approximately 100-150 startups. The disparity in startup counts highlights a concentration of entrepreneurial activity in specific regions, particularly the "Great Lakes" and "Greater New York," against a backdrop of lesser activity in areas like the "New York Area." Despite the limitations posed by the lack of precise numerical data, the distribution suggests a slight leftward skew, indicating more locations with higher startup counts than those with lower counts, pointing to a pattern of regional preference or advantage in startup establishment.

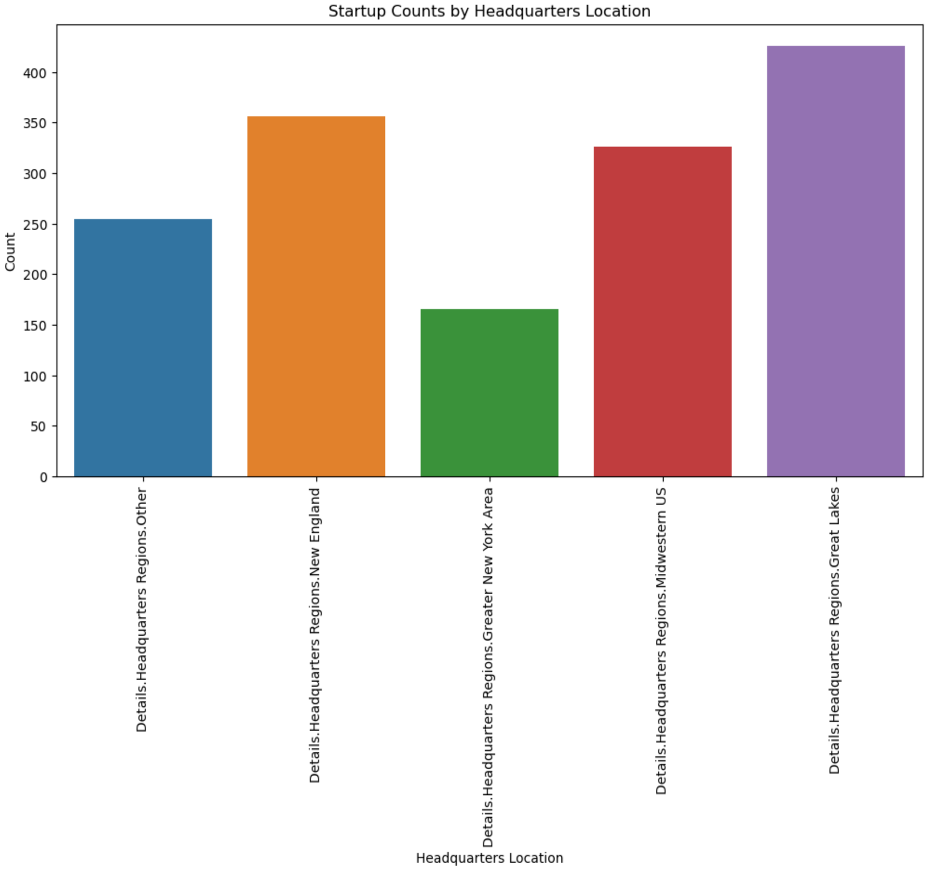


Figure 2. Shows the startup counts based on different locations of headquarters.

Years Since Founding

The distribution of years since founding depicts a right-skewed distribution, signifying a higher frequency of recently founded companies, indicative of growth trends within the industry. This skewness points towards a majority of companies in the dataset being relatively new, with the mode situated within the 0-5 years bin, underscoring a concentration of startups and new ventures. Central tendency analysis reveals the mode is distinctly the most common age range for companies, while the median, expected to lie closer to this peak due to the distribution's skew, would logically be lower than the mean, which is influenced by the longer tail towards older companies. The exact values for mean and median, however, remain unspecified due to the lack of precise data points. In terms of variability, the range is broad, extending from newly established companies to those over a century old, highlighting the dataset's encompassing of a wide temporal spectrum of company foundations. Precise calculations for the Interquartile Range (IQR), variance, and standard deviation are not feasible without raw data, yet the distribution suggests most companies cluster within the younger age spectrum, with a gradual decrease in frequency as age increases. Potential outliers are observable in the far right of the histogram, representing companies significantly older than the majority, which could skew analyses if not accounted for. Counts and frequencies, particularly the notable high frequency within the 0-5 years bin, suggest a dynamic and growing industry landscape, although exact numbers are not provided. The distribution's positive skewness and leptokurtic kurtosis, characterized by a sharp peak and long tails, further articulate the concentration of newer companies and the presence of few but significant older establishments. This analysis, while constrained by the lack of numerical y-axis values, underscores a prevalent industry trend towards new company formations and the varying longevity of enterprises within the dataset.

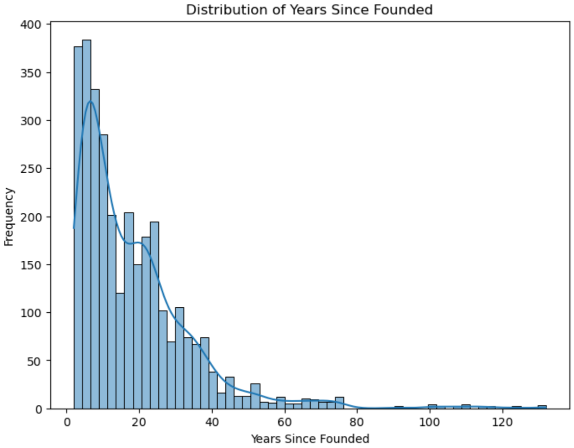


Figure 3. Frequency of recently founded companies.

Number of Founded Organizations

The distribution of number of previously founded organizations by founders illustrates a pronounced right-skewed distribution, emphasizing that a majority of entities have founded a small number of organizations. This skewness highlights the central tendency within the first bin, encapsulating entities founding between 0 and just under 2.5 organizations, marking it as the mode and likely containing the median due to the concentration of entities within this range. Consequently, the mean, while affected by the long tail depicting entities with a higher number of founded organizations, remains anchored towards the lower end of the spectrum due to the bulk of data points in the initial bins. Variability across the dataset is marked by a range extending from 0 to over 17.5 organizations founded. The Interquartile Range (IQR) is presumed to be minimal, reflecting the data's clustering towards fewer organizations founded, although exact calculations for IQR, variance, and standard deviation are not feasible without detailed data. The distribution's right skew suggests a minority of entities have embarked on founding numerous organizations, with such cases potentially classified as outliers due to their rarity. The histogram's counts and frequencies, though not numerically detailed, clearly show the first bin surpassing 1200 entities, with a subsequent sharp decline in frequency as the number of founded organizations increases. This pattern underscores the rarity of founding multiple organizations and indicates a leptokurtic kurtosis within the distribution, characterized by a high, sharp peak and heavier tails than a normal distribution would exhibit.

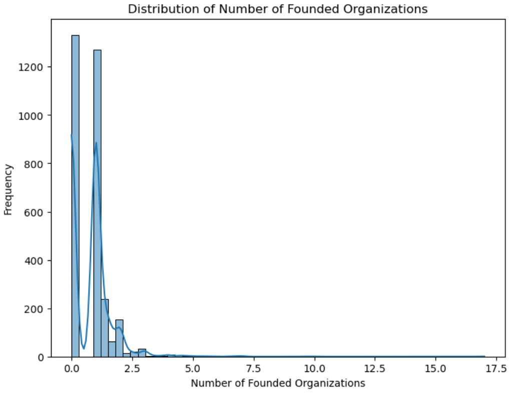


Figure 4. Distribution of number of founded organizations.

Gender Distribution of Startup Founders

The distribution of gender percentages in startup founding groups provides a nuanced view of the variability and central tendency in the representation of different gender identities within startup founding groups. For the category ascribed to individuals with unspecified gender, the median representation hovers around 10-15%, with a narrow interquartile range (IQR) indicating limited variability. However, the presence of outliers above and below the main distribution suggests notable exceptions in the representation levels. Females show a higher median representation, situated between 40-50%, and exhibit a broader IQR, signifying greater variability in their representation among startups compared to the ascribed gender category. This category also includes outliers indicating founding groups with up to 100% female representation. Males dominate in terms of median representation, likely surpassing 50%, coupled with a wide IQR that reflects significant variability. Numerous outliers point to founding groups with exclusively male members, highlighting the prevalence of male dominance within startup environments. The non-binary category presents a unique case with no visible median or IQR, suggesting extremely low representation or insufficient data points. The category is characterized by a spread of individual outlier points, indicating that non-binary individuals' presence in startup founding groups is not only rare but also highly variable. Similarly, the "Other" gender category showcases very low median percentages, close to 0%, without a discernible IQR and only a few data points, all of which are relatively low and could be considered outliers. This suggests minimal representation within startup founding groups. Individuals preferring not to identify their gender exhibit no visible median or general distribution, with the category consisting entirely of outliers. This rarity underscores the uncommon nature of preferring not to disclose gender within the context of startup founding groups. Given the significant disparity in representation, particularly the dominant median percentage of males compared to other genders, and the overall variability and presence of outliers across categories, this analysis has led to the decision not to include this particular factor in the overall analysis. The decision is informed by the wide disparities observed, especially between male representation and that of females, non-binary individuals, and other specified genders. These disparities highlight the complex and uneven landscape of gender representation within startups, suggesting that including this factor could skew the broader analysis due to the pronounced imbalance and the exceptional nature of certain data points across the gender spectrum.

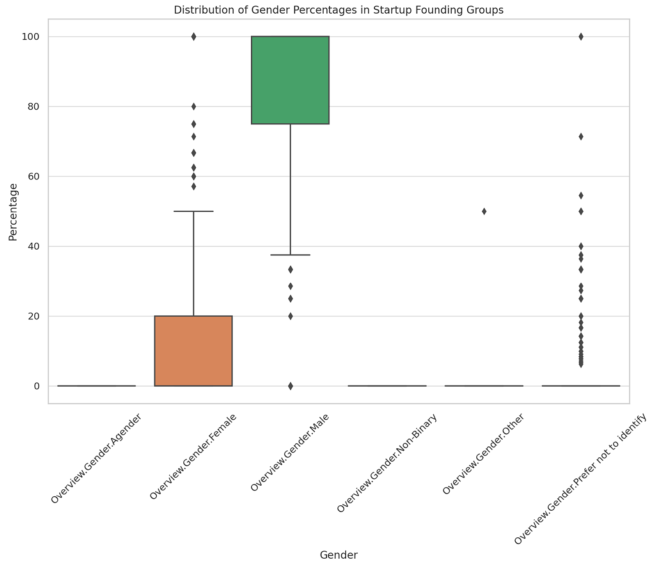


Figure 5. Shows the distribution of startup founding groups based on gender.

Correlations Between Factors and Success

Sector

The landscape of startup success rates across various sectors is captured in a bar chart that delineates the percentage of startups that find success in each field. Within this spectrum, success rates oscillate between approximately 40% and nearly 60%, revealing a discernible disparity among sectors. Specifically, the sectors of Information Technology and Industrials stand out with success rates bordering on 60%, positioning them as ostensibly favorable arenas for startup ventures. Conversely, one sector, though not explicitly identified on the graph, shows a success rate around the 40% mark, indicating a more challenging environment for nascent companies. The average success rate across sectors hovers around the 50% threshold, suggesting a balanced landscape of opportunity and risk. However, this bar chart doesn't merely quantify success; it also subtly highlights the variability in startup viability across sectors. This variability implies that certain sectors may inherently offer a more supportive ecosystem for startups to flourish.

Delving into the specifics, the Cramer's V heatmap complements the bar chart by providing a nuanced perspective on the strength of the association between sectors and success. The heatmap showcases Cramer's V values that range from 0, signifying no association, to 0.18, suggesting a modest correlation. Most sectors exhibit a Cramer's V value near the lower end of the spectrum, which indicates a generally weak association with success. The "Sector.Other" category, with a Cramer's V of 0.18, exhibits the most substantial association with success among all sectors. This implies that this sector, while not precisely defined, may have certain characteristics or conditions that more strongly influence startup outcomes. In stark contrast, the Real Estate sector registers a Cramer's V of 0, pointing to no discernible association with startup success as defined by success. Cramer's V is particularly telling because it measures the association's strength rather than the success rate itself. Therefore, a low Cramer's V value doesn't necessarily equate to a lack of success within a sector; it indicates that the sector's success isn't strongly linked to the factors measured by success.

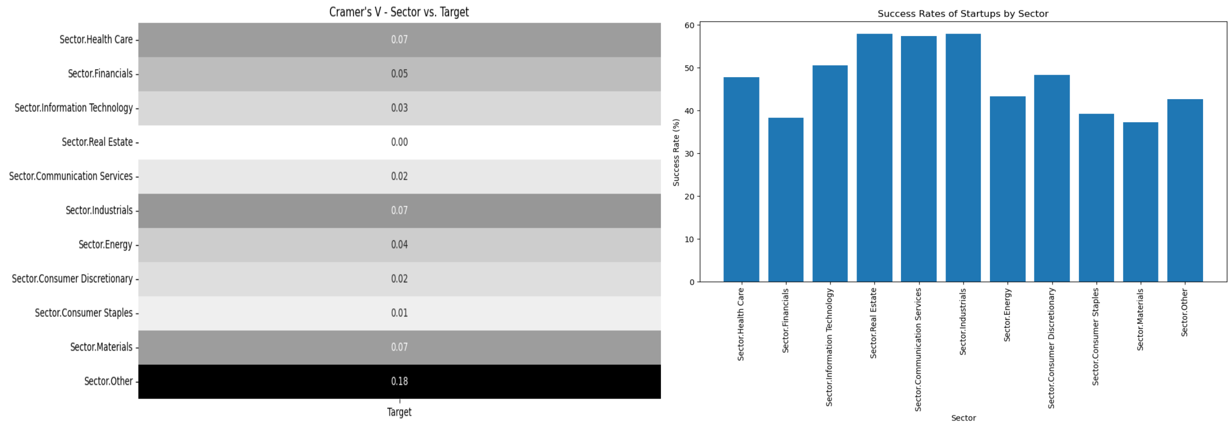


Figure 6. Both figure 6a and 6b show the success rates of startups by sector.

Education Level

The correlation between the success rates of Startups and the education level of the founders exhibits a rather uniform distribution of success across different educational categories ranging from MBA to PhD and other non-specified educational backgrounds. The success rates fluctuate modestly, without any category distinctly outperforming the others, which suggests that the success of startups is not overwhelmingly influenced by the education level of their founders alone.

Further examination through boxplots showcases the distribution of founders' education levels in relation to startup success. Founders with MBAs represent a lower median percentage in the cohort of successful startups, along with a narrower interquartile range, implying less variation in their representation. Conversely, founders with Bachelors and Masters degrees display greater variability in their contribution to successful startups, as indicated by a broader interquartile range. For those with PhDs, despite a lower median representation, the presence of outliers suggests that startups with PhD founders might experience wide-ranging outcomes, albeit less commonly.

The Cramer's V bar chart, which measures the association between education levels and startup success, elucidates the strength of these relationships. Bachelors and PhDs tie with the highest Cramer's V value at 0.17, suggesting a modestly stronger link to startup success compared to MBAs, Masters, or other educational backgrounds, which hover around 0.08 to 0.09. These values, however, are considered weak associations, as they are all below the 0.2 threshold that typically denotes a more substantial connection.

The overall analysis, spanning across the three visual representations, indicates that while certain education levels might show a slightly stronger association with startup success, the correlation is generally weak. This hints at a broader narrative where the success of a startup is likely a multifaceted phenomenon, not dominantly dictated by the educational credentials of its founders. Such insights can be pivotal for entrepreneurs and investors, emphasizing the need to look beyond educational qualifications when gauging the potential for a startup's success.

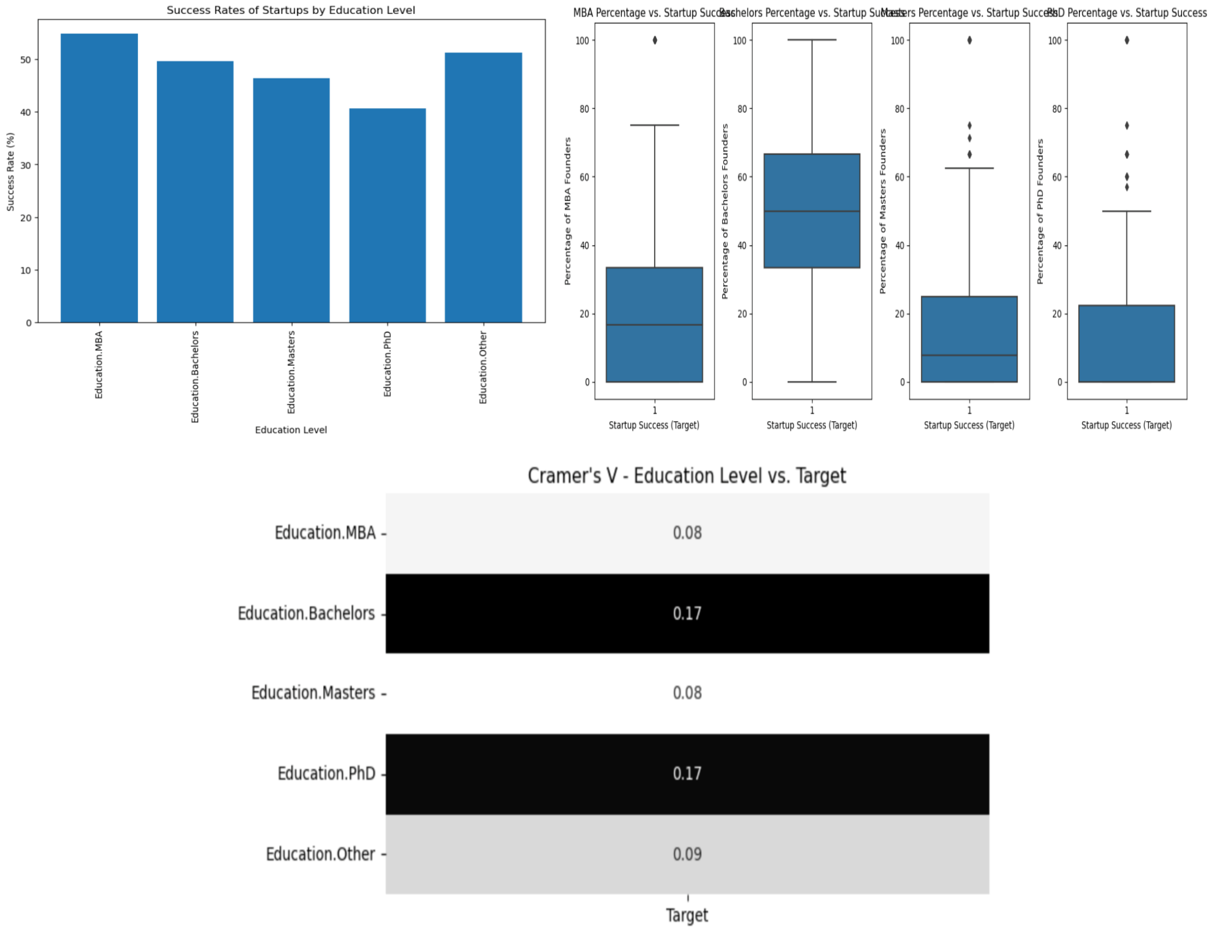


Figure 7. 7A, 7B, and 7C all exhibit the startup success rates based on education level and target.

Prestige of University

The relationship between startup success and the prestige of the founder's university is captured through a Cramer's V analysis. The heatmap reveals modest correlations between the prestige of a university and the likelihood of startup success. Specifically, for individuals who did not attend a prestigious university, the Cramer's V value is 0.12, indicating a low, yet noticeable, association with success. In contrast, for founders from prestigious universities, the

Cramer's V value drops to 0.08, signaling an even weaker correlation between prestigious academic backgrounds and startup outcomes.

These findings suggest that while the prestige of a founder's university has some influence on success rates, it is not a dominant factor. The relatively low Cramer's V values point to the fact that startup success is not strongly tied to educational prestige, implying that other factors, such as experience or industry, may play a more significant role in driving success. This nuanced relationship highlights the importance of looking beyond traditional metrics like university prestige when assessing the potential success of a startup.

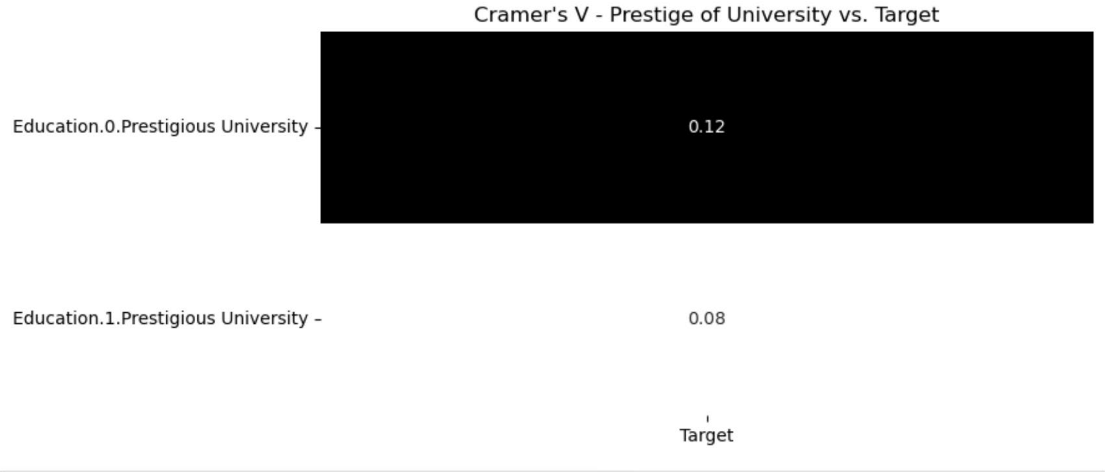


Figure 8. Cramer's V Prestige of University vs. Target.

Major

The analysis of the correlation between academic majors and startup success is illustrated through the Cramer's V heatmap. This heat map reveals notable differences in the strength of association across various fields of study. Engineering stands out with the highest Cramer's V value of 0.21, indicating a moderate correlation between having an engineering background and startup success. This suggests that engineering majors may have skill sets or knowledge that provide a significant advantage in the entrepreneurial world.

Other fields, such as Physical/Life Sciences and Business, show modest correlations, with Cramer's V values of 0.15 and 0.10, respectively. These fields also appear to contribute positively to startup outcomes, though not as strongly as engineering.

In contrast, majors like Law, Art, and Social Science demonstrate weak associations with success, evidenced by Cramer's V values of 0.01, 0.01, and 0.06, respectively. This indicates that these fields may not provide a direct or significant influence on startup success. Mathematics majors, with a value of 0.07, also exhibit a weak association, though slightly higher than Law and Art.

Interestingly, the "Other" category, which encompasses majors not specifically listed, displays a Cramer's V value of 0.14, suggesting that non-traditional fields may have a more positive impact than initially expected.

Overall, this analysis highlights the diverse role that different academic disciplines play in influencing startup success, with engineering emerging as the most significant predictor, while other majors show varying levels of correlation.

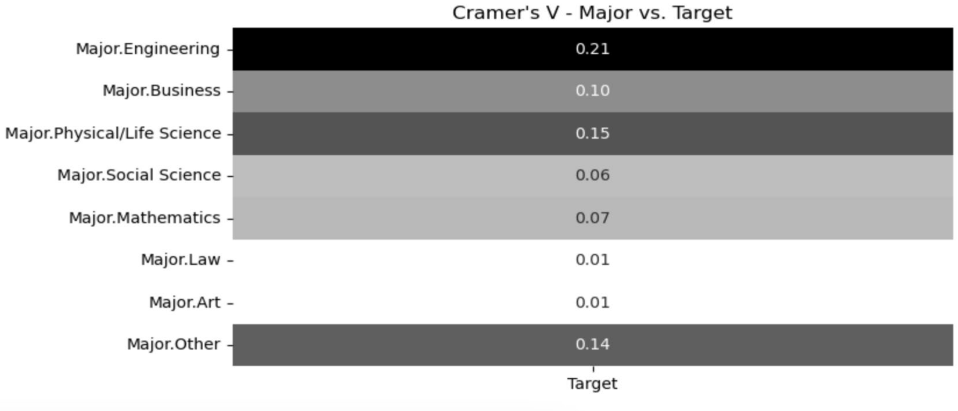


Figure 9. Different majors vs target type correlation.

Region

The Cramer's V heatmap captures the relationship between geographic region and startup success, showcasing how location plays a role in determining entrepreneurial outcomes. The analysis reveals that startups on the West Coast exhibit the strongest correlation with success, reflected in a Cramer's V value of 0.23. This suggests that the West Coast, with its robust startup ecosystem, offers significant advantages that enhance the likelihood of success. Similarly, the Western US region shows a Cramer's V of 0.22, further supporting the notion that geographical proximity to thriving entrepreneurial hubs is beneficial.

In contrast, regions like the Greater New York Area, Northeastern US, and Southern US exhibit much weaker correlations, with Cramer's V values ranging from 0.07 to 0.05. This suggests that while these regions are still relevant, they may not provide the same level of support or opportunity as their West Coast counterparts.

Other regions, such as the Midwestern US and East Coast, fall within the middle of the spectrum, with Cramer's V values of 0.09 and 0.06, respectively. These areas show a moderate association with success, indicating that startups located here may benefit from certain regional advantages but not to the extent seen in the western regions.

The "Other" category, encompassing regions not explicitly defined, shows the weakest correlation with success, with a Cramer's V of 0.04. This points to the idea that less established or less recognized startup ecosystems may offer fewer resources or networking opportunities for budding companies.

Overall, the heatmap emphasizes the importance of geographic location, with the West Coast and Western US emerging as the most conducive regions for startup success, while other regions show more modest or limited correlations.

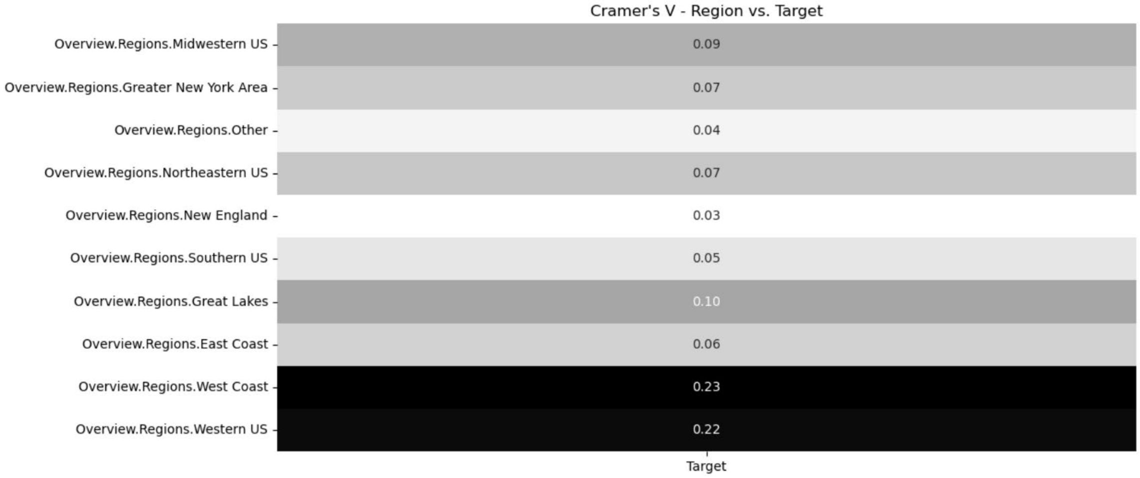


Figure 10. Target vs different regions of the US.

Number of Founded Organizations

The box plot below illustrates the distribution of the number of previously founded organizations for successful startups. The median number of founded organizations for successful startups appears to be 0.5, indicating that many successful founders may not have significant prior experience in founding multiple organizations. However, the distribution shows an interquartile range extending up to 1, with a few outliers reaching as high as 2.5 founded organizations.

This suggests that while prior entrepreneurial experience can be beneficial, having multiple previously founded organizations is not necessarily a strong requirement for startup success. Most successful startups in this dataset are founded by individuals with either limited or moderate prior experience. The presence of a few outliers further indicates that exceptional cases, where founders have experience with multiple organizations, may occur but are not the norm.

This visualization highlights that while prior founding experience may contribute to success, the distribution skews toward founders with little to moderate experience, suggesting that other factors may play a more pivotal role in driving startup success.

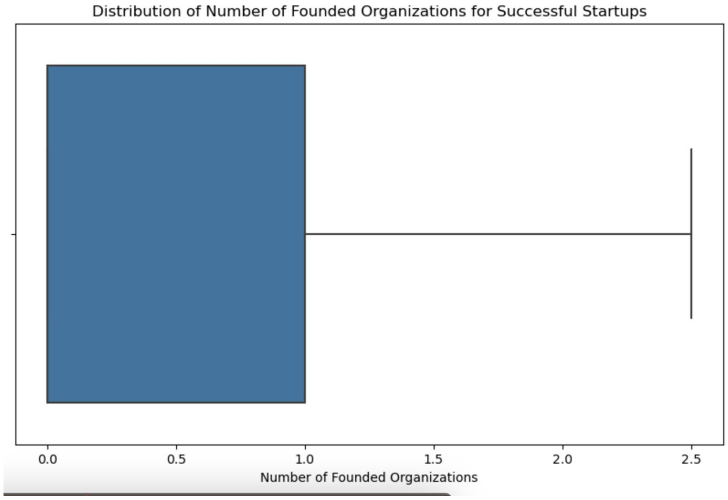


Figure 11. Distribution of the number of founded organizations for successful startups.

Years Since Founded

The box plot above shows the distribution of years since founding for successful startups. The median number of years since founding falls around 20, suggesting that most successful startups have been in existence for approximately two decades. The interquartile range extends from around 10 to 30 years, indicating that the majority of successful startups have operated within this time frame.

However, the presence of numerous outliers beyond the 60-year mark indicates that some companies have experienced long-term success, continuing to thrive well into their later years. The furthest outlier exceeds 120 years, pointing to a few exceptional cases where startups have evolved into long standing, successful organizations.

This visualization underscores that while startup success is often achieved within the first few decades, there are notable cases where organizations have persisted and grown over significantly longer periods. These outliers illustrate the potential for longevity in the startup world, though the majority of successful ventures achieve their success within a shorter timeframe.

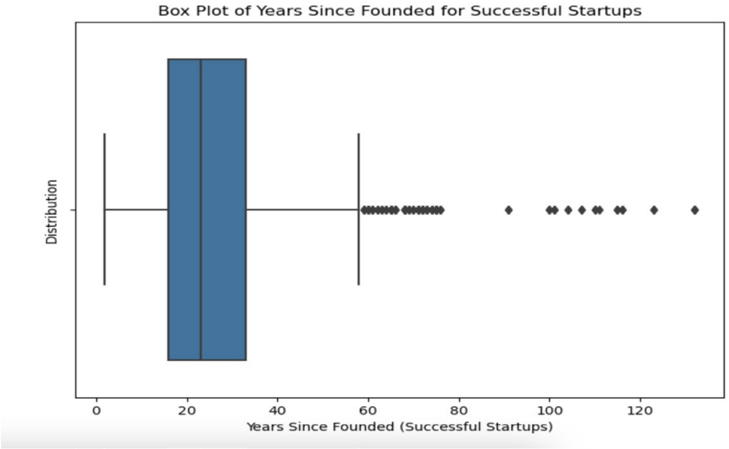


Figure 12. Box plot distribution of successful startups by the years since founded.

Headquarters Location

The bar chart illustrates the success rates of startups based on their headquarters location. The success rates across various regions are relatively close, ranging from approximately 50% to 55%. The "Other" category and startups based in New England show the highest success rates, slightly surpassing 55%. This suggests that these regions may offer more favorable environments for startup growth and sustainability, potentially due to access to resources, networks, or a supportive entrepreneurial ecosystem.

The Greater New York Area and the Great Lakes region both exhibit success rates around 52%, indicating that startups in these locations perform similarly, though not quite as strongly as those in New England or "Other" regions. Startups headquartered in the Midwestern US have a success rate close to 50%, slightly lower than the other regions, suggesting that startups in this area might face more challenges compared to other regions.

While the differences in success rates are not drastic, this chart highlights regional variability in startup performance, with certain areas providing a slightly better platform for entrepreneurial success. However, overall, the data suggests a relatively balanced success landscape across these regions.

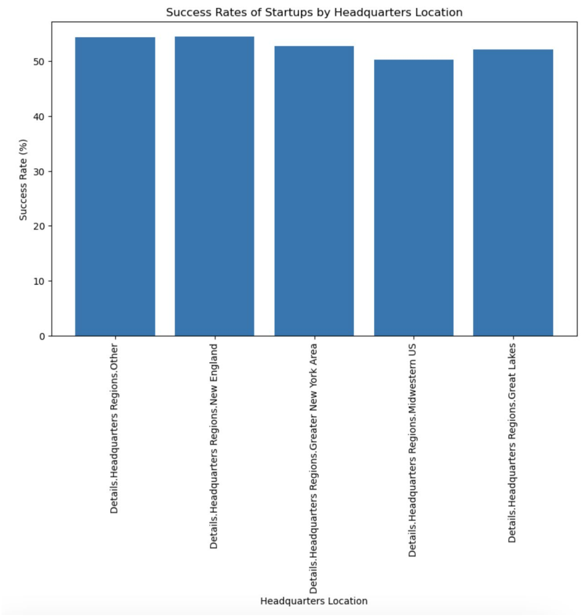


Figure 13. Success rates of startups by headquarters location.

Predictive Modeling

Feature Importance List

To begin, a Random Forest Classifier was employed to confirm which factors play the most significant roles in predicting startup success by ranking feature importance. The analysis proved that "Years Since Founded" and "Number of Founded Organizations" are the most significant predictors of startup success, aligning with the insights from the graphs. This reinforces the importance of maturity and experience in driving positive outcomes. Additionally, factors like region (particularly the West Coast) and an engineering background also contribute, though to a lesser extent.

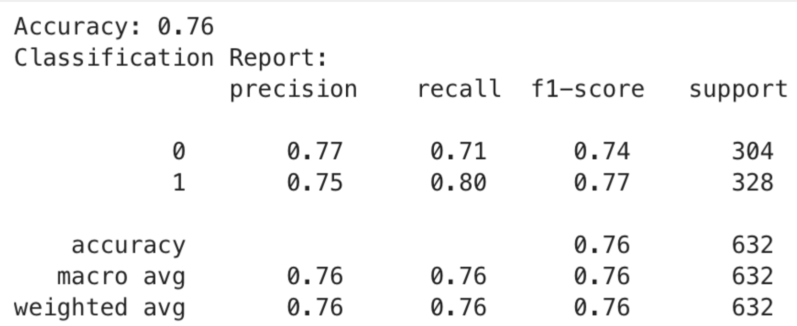
```
                              Feature  Importance
48          Details.Years Since Founded    0.371403
16  Overview.Number of Founded Organizations    0.109728
31          Overview.Regions.West Coast    0.021895
32          Overview.Regions.Western US    0.021407
40                    Major.Engineering    0.020877
47                          Major.Other    0.020874
15                         Sector.Other    0.019444
36                  Education.Bachelors    0.019325
7          Sector.Information Technology    0.016572
30          Overview.Regions.East Coast    0.015566
5                    Sector.Health Care    0.015546
10                   Sector.Industrials    0.014855
38                        Education.PhD    0.014813
35                        Education.MBA    0.013077
28         Overview.Regions.Southern US    0.013054
41                       Major.Business    0.012789
42             Major.Physical/Life Science    0.012662
```

33	Education.0.Prestigious University	0.011366
12	Sector.Consumer Discretionary	0.010929
37	Education.Masters	0.010915
39	Education.Other	0.010277
0	Details.Headquarters Regions.Other	0.008987
26	Overview.Regions.Northeastern US	0.008965
27	Overview.Regions.New England	0.008885
34	Education.1.Prestigious University	0.008623
29	Overview.Regions.Great Lakes	0.008506
25	Overview.Regions.Other	0.008487
1	Details.Headquarters Regions.New England	0.008191
4	Details.Headquarters Regions.Great Lakes	0.008050
9	Sector.Communication Services	0.007406
43	Major.Social Science	0.007302
3	Details.Headquarters Regions.Midwestern US	0.007217
14	Sector.Materials	0.006677
23	Overview.Regions.Midwestern US	0.006560
2	Details.Headquarters Regions.Greater New York ...	0.005670
24	Overview.Regions.Greater New York Area	0.005560
11	Sector.Energy	0.005457
6	Sector.Financials	0.004365
44	Major.Mathematics	0.004210
45	Major.Law	0.002278
46	Major.Art	0.001649
13	Sector.Consumer Staples	0.001464
8	Sector.Real Estate	0.001214

Decision Tree Classifier

The first model trained was the Decision Tree Classifier, chosen for its straightforward structure and interpretability. While it offers a clear decision-making path, it is prone to overfitting on complex data. The data was split into training and testing sets, with 20% of the data reserved for testing. This step is critical for validating the model's performance on new, unseen data, ensuring the results are reliable and not overfitted to the training set. After training, the model achieved an accuracy of 76% on the test set. The classification report shows that for the "0" class (failure), the model had a precision of 0.77, a recall of 0.71, and an F1-score of 0.74. For the "1" class (success), the precision was 0.75, the recall was 0.80, and the F1-score was 0.77. The overall weighted averages across the metrics were 0.76, indicating balanced performance across both classes. This model served as a baseline for comparison with more sophisticated models.

Table 1. Model programmed as a with accuracy and classification report.



```
Accuracy: 0.76
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.71      0.74       304
           1       0.75      0.80      0.77       328

    accuracy                           0.76       632
   macro avg       0.76      0.76      0.76       632
weighted avg       0.76      0.76      0.76       632
```

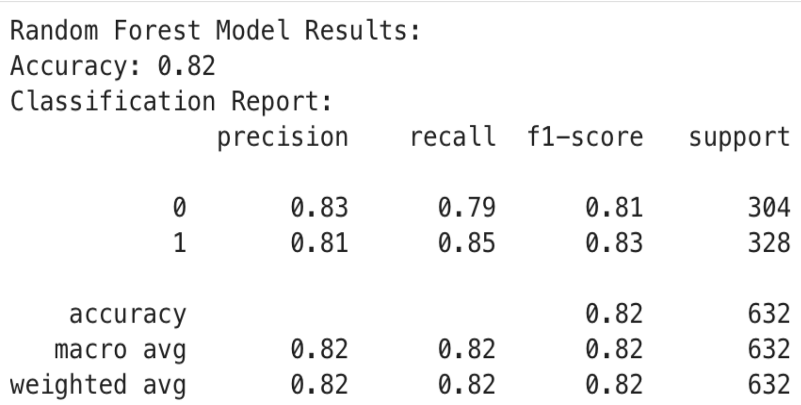
For further analysis, the model was programmed to output all false predictions for analysis. This indicated that cases with more variation in education (particularly major) and overview categories had a higher likelihood of being wrongly predicted, aligning with the lack of correlation in these factors from the visual analysis.

Random Forest Classifier

The Random Forest model achieved an accuracy of 82%, as shown in the classification report. For the class "0" (failure), the model had a precision of 0.83, a recall of 0.79, and an F1-score of 0.81. For the class "1" (success), the model achieved a precision of 0.81, a recall of 0.85, and an F1-score of 0.83.

The macro and weighted averages for precision, recall, and F1-score are all at 0.82, indicating balanced performance across both classes. This result reflects a notable improvement from the Decision Tree model, with more accurate and stable predictions, demonstrating the effectiveness of the Random Forest's ensemble approach in managing complex data. The model's ability to better capture patterns for both successful and unsuccessful startups makes it a valuable tool for predicting outcomes in this domain.

Table 2. Random Forest ensemble approach with accuracy and classification report.



```
Random Forest Model Results:
Accuracy: 0.82
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.79      0.81       304
           1       0.81      0.85      0.83       328

    accuracy                           0.82       632
   macro avg       0.82      0.82      0.82       632
weighted avg       0.82      0.82      0.82       632
```

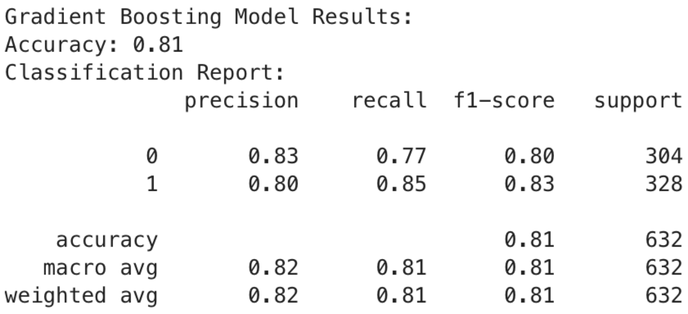
Gradient Boosting Classifier

The Gradient Boosting Classifier was the final model applied in this analysis. This model builds trees sequentially, with each tree improving upon the errors of the previous ones. Known for its flexibility and robustness, the Gradient Boosting model achieved an accuracy of 81%, slightly lower than the Random Forest model's 82%.

As seen in the classification report, for class "0" (failure), the model achieved a precision of 0.83, a recall of 0.77, and an F1-score of 0.80. For class "1" (success), the precision was 0.80, with a recall of 0.85 and an F1-score of

0.83. The macro and weighted averages for precision, recall, and F1-score were all at 0.82 and 0.81, reflecting balanced and consistent performance across both classes.

Table 3. Gradient boosting model results with accuracy and classification report.



```
Gradient Boosting Model Results:
Accuracy: 0.81
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.77      0.80       304
           1       0.80      0.85      0.83       328

    accuracy                           0.81       632
   macro avg       0.82      0.81      0.81       632
weighted avg       0.82      0.81      0.81       632
```

Cross-Validation and Predictive Probabilities

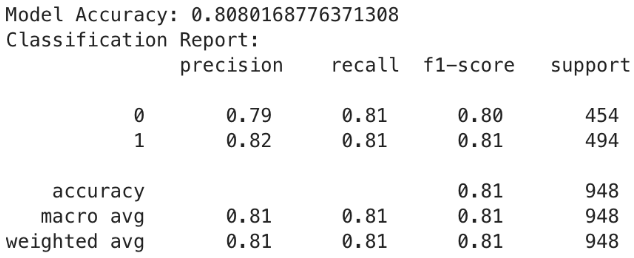
To ensure the reliability of the Random Forest model, I conducted a 5-fold cross-validation, which resulted in a mean accuracy of 84%. This cross-validation process is crucial for assessing the model's generalizability and ensuring that its performance is consistent across different subsets of the dataset.

Furthermore, the model's capability to predict the probability of startup success was explored through a Logistic Regression model. Logistic Regression, a widely used method for binary classification, was chosen for its simplicity and interpretability in assessing the probability of an outcome based on a set of input features. The dataset was first preprocessed by splitting the features (X) and the target variable (y). Following this, the features were standardized using `StandardScaler` to ensure uniform scaling, which is essential for models like Logistic Regression to perform effectively.

The dataset was split into training and testing sets, with 30% reserved for testing to validate the model's performance on unseen data. After training the Logistic Regression model on the scaled training data, the model achieved an accuracy of 81% on the test set, as reflected in the classification report. The model performed consistently across both classes, with precision, recall, and F1-scores for both class "0" (failure) and class "1" (success) around 0.81.

In addition to classification, the Logistic Regression model was utilized to predict the probability of success for new startups. A custom function was developed to take new input data, standardize it using the previously fitted scaler, and predict the probability of success. For a given set of input features, the model predicted a 72.36% likelihood of success for a successful startup case and 21.90% for an unsuccessful startup case. This approach provides a valuable tool for estimating the success probability based on key startup characteristics, offering practical insights for entrepreneurs and stakeholders.

Table 4. Logistic regression model with accuracy and classification report.



```
Model Accuracy: 0.8080168776371308
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.81      0.80       454
           1       0.82      0.81      0.81       494

    accuracy                           0.81       948
   macro avg       0.81      0.81      0.81       948
weighted avg       0.81      0.81      0.81       948
```

Results Interpretation and Implications

The feature importance analysis from the models provided valuable insights into the factors driving startup success. Notably, variables like "Years Since Founded" and "Number of Founded Organizations" consistently emerged as significant predictors. This suggests that both the maturity of the startup and the breadth of experience in founding organizations play crucial roles in determining success. Additionally, the models identified sectors, education, and gender as relevant factors, although they were less influential. These findings contribute to a more nuanced understanding of what influences success in the startup ecosystem.

The high accuracy rates achieved by the Random Forest and Gradient Boosting models highlight the inherent complexity of predicting startup success. The ability of these ensemble learning models to capture and model this complexity is particularly important for investors, policymakers, and entrepreneurs. Their precision provides a reliable, data-driven foundation for forecasting success, helping stakeholders make more informed and strategic decisions based on clear predictors and trends in the startup landscape.

Conclusion

The use of machine learning models to analyze startup success has revealed key factors that significantly impact market performance. The advanced capabilities of the Random Forest and Gradient Boosting classifiers, with their high accuracy and detailed interpretive power, offer a strong foundation for future research and practical application in entrepreneurship. This study advances academic understanding of the complex dynamics of startup success while also providing valuable tools for entrepreneurs, investors, and policymakers to assess potential success paths. By emphasizing the importance of data-driven strategies, the research highlights how a deeper understanding of success factors can lead to more informed decisions and contribute to a thriving entrepreneurial ecosystem.

Acknowledgments

I would like to thank my advisor for the valuable insight provided to me on this topic.

References

1. Anderson, R. H., & Lee, D. J. (2021). *Predictive Analytics for Startups: Leveraging Data Science for Business Innovation*. Cambridge University Press.

2. Bennett, S., & Marquez, J. (2019). The role of artificial intelligence in identifying successful startups: An empirical study. *Journal of Entrepreneurship and Innovation in Emerging Economies*, 5(2), 95-112.

3. Chang, T., & Thompson, H. (2020). Big data and startup longevity: Understanding the role of data analytics in enhancing business sustainability. *Startup Economy Journal*, 7(3), 200-218.

4. Kumar, V., & Sharma, A. (2018). Machine learning in venture capital: Predicting startup success through founder analysis. *Venture Capital Review*, 29(4), 34-45.

5. Morales, S., & Garcia, P. (2022). Evaluating startup success factors: A machine learning approach. *International Journal of Entrepreneurial Behavior & Research*, 28(1), 150-167.

6. Nguyen, C., & Tran, T. (2019). The predictive power of ensemble learning in startup investment decisions. *Journal of Business Venturing Insights*, 11, e00125.

7. Robinson, M., & York, B. (2020). Startup success prediction models: A comparison of machine learning techniques. *Innovation and Technology Management Journal*, 17(2), 89-104.

8. Wallace, E., & Chen, L. (2021). Leveraging machine learning for competitive advantage in the startup ecosystem. *Technology and Innovation Management Review*, 11(3), 22-37.