Unravelling the Dynamics of Startup Success Prediction: A Thesis on the Comparative Study of Machine Learning Models and Techniques

Ankur Napa

Research Proposal

April 2023

# Abstract

In the rapidly evolving and unpredictable landscape of startup ventures, the ability to accurately predict their prospects of success is of paramount importance to a diverse group of stakeholders - investors, startup founders, and policy regulators. The primary objective of this research endeavor is to explore and untangle the intricate facets of forecasting startup outcomes by implementing a wide array of machine learning models and methodologies.

To provide a solid foundation for our investigation, a meticulous review of the available literature will be undertaken. This will help us to recognize the key drivers of startup success, and to gauge the current use of machine learning in this context.

Our data source for this study is Crunchbase, an extensive dataset rich with relevant features and indicators of success. The information gleaned from Crunchbase will serve as the ground for training and testing a variety of machine learning models. In this investigation, our primary focus will be centered around models such as logistic regression, decision trees, random forests, and XGBoost.

We will evaluate the effectiveness of these models based on traditional performance metrics including accuracy, precision, recall, F1 score, and the area under the receiver operating characteristic (ROC) curve. With a comprehensive comparative analysis, our research strives to pinpoint the most suitable machine learning models and methodologies for accurately predicting the success of startups.

The insights gathered from this study will provide valuable directions for future academic inquiries and practical implementations in the field of startup success prediction using machine learning. Notably, our findings will be grounded on the Crunchbase dataset, ensuring a robust and reliable basis for our conclusions.

# Table of Contents

[**Abstract** 3](#_Toc132543181)

[**1. Background** 5](#_Toc132543182)

[**2. Related work** 5](#_Toc132543183)

[**3.Research Questions** 5](#_Toc132543184)

[**4. Aim & Objectives** 5](#_Toc132543185)

[**5. Significance of Study** 5](#_Toc132543186)

[**6. Scope of Study** 5](#_Toc132543187)

[**6.1 In scope**: 5](#_Toc132543188)

[**6.2 Out of scope:** 5](#_Toc132543189)

[**6.3 Reason for defining the scope:** 5](#_Toc132543190)

[**7. Research Methodology** 5](#_Toc132543191)

[**7.1 Introduction** 5](#_Toc132543192)

[**7.2 Business Understanding:** 5](#_Toc132543193)

[**7.6 Why these models ?** 6](#_Toc132543194)

[**Evaluation Matrix:** 6](#_Toc132543195)

[**8. Required Resources** 7](#_Toc132543196)

[**8.1 Hardware requirements:** 7](#_Toc132543197)

[**8.2 Software requirements:** 7](#_Toc132543198)

[**9. Research Plan** 7](#_Toc132543199)

[**10. References** 8](#_Toc132543200)

**List of Figures**

[Figure 1 : Ten Stage fermentation process 9](#_Toc124626669)

[Figure 2: Fermentation Process 10](#_Toc124626670)

[Figure 3: Our Research Focus on these colored processes 10](#_Toc124626671)

[Figure 4 : Process Flow 12](#_Toc124626672)

# 1. Background

The significance of startups in the global economy is immense, as they play a crucial role in creating employment, driving innovation, and attracting investments. By offering diverse workforce opportunities and introducing groundbreaking products and services, startups boost local and regional economies, enhance productivity, and elevate living standards. As they develop and broaden their reach, startups contribute to economic expansion and diversification, making economies more robust and well-rounded. Furthermore, by nurturing innovation and entrepreneurial attitudes, startups bolster a nation's standing in the global market, securing investments, forging international partnerships, and sustaining economic growth.

Startups also focus on addressing societal, ecological, and economic challenges, leading to a positive social impact while generating profits. This dual emphasis fosters new avenues for growth and collaboration. Thriving startups can catalyse the development of new ecosystems that support and encourage other entrepreneurial initiatives, creating an environment where entrepreneurs can succeed. This interplay between startups and their supporting ecosystems is instrumental in the overall development and sustainability of the global economy, propelling innovation, and progress well into the future.

The need for accurate prediction of startup success is vital for various stakeholders involved in the entrepreneurial ecosystem. This includes investors, entrepreneurs, policymakers, and support organizations. Accurate prediction of startup success can lead to more informed decision-making, optimized resource allocation, and overall better outcomes for everyone involved.

For investors, the ability to predict startup success is crucial in identifying potential high-growth ventures and making sound investment decisions. By allocating funds to startups with a higher likelihood of success, investors can maximize their returns and reduce the risk associated with their investments. This also ensures that capital is channelled towards the most promising ventures, accelerating innovation and economic growth.

Entrepreneurs benefit from accurate success predictions as it enables them to identify their strengths and weaknesses and make necessary adjustments to their strategies. By understanding the factors that contribute to their success, entrepreneurs can focus on areas that need improvement, increasing their chances of survival and growth. This can lead to more sustainable business models and a higher likelihood of creating a lasting impact in their respective markets.

For policymakers, accurate startup success prediction is essential for designing and implementing effective policies and support programs that foster a thriving entrepreneurial ecosystem. By understanding the factors that contribute to startup success, policymakers can target their efforts and resources towards initiatives that have the highest potential for stimulating innovation, job creation, and economic development. This can result in more efficient use of public funds and a higher return on investment for the community.

Support organizations, such as accelerators, incubators, and mentoring programs, also benefit from accurate startup success prediction. By identifying the key factors that drive success, these organizations can tailor their programs to address the specific needs of startups and provide targeted support. This allows them to optimize their resources, improve their program's efficacy, and enhance the overall impact on the startup ecosystem.

In summary, the need for accurate prediction of startup success is paramount for informed decision-making and optimizing resources across various stakeholders in the entrepreneurial ecosystem. By understanding the factors that contribute to success, investors, entrepreneurs, policymakers, and support organizations can work together to foster a more robust and thriving startup landscape, driving innovation, job creation, and economic growth.

# 2. Related work

## Problem Statement:

Predicting the success of startups presents numerous obstacles, one of which is their inherently dynamic nature. Startups continually adapt due to factors such as swift technological progress, fluctuating market conditions, and competitive forces, making it challenging for static models to accurately assess their potential.

Data quality and accessibility are also critical concerns when predicting startup success. Reliable predictions depend on comprehensive and accurate data; however, many startups lack an extensive historical record. Furthermore, the data that is available may be limited, incomplete, or biased, adding to the intricacy of the modelling process.

The diverse nature of startups contributes to the difficulties in predicting their success. Given the wide range of business models, industries, target markets, and developmental stages, constructing generalizable models that accurately forecast success across various startups becomes a daunting endeavor.

Finally, the subjective nature of defining success adds to the challenges of predicting startup outcomes. Stakeholders may prioritize different criteria, such as financial metrics like revenue or profitability, or emphasize social impact or market disruption. The lack of a universally accepted definition of success increases the complexity of developing effective predictive models.

1.2.2 Related Work

In the rapidly evolving landscape of startups, predicting their success or failure has become a critical aspect of entrepreneurial strategy and investment decision-making. A considerable amount of research has been conducted in this area using machine learning models, demonstrating their capacity to effectively predict startup outcomes.

(Sadatrasoul et al., 2020) made a significant contribution to this field by developing a business success failure (S/F) prediction model for Iranian startups. They conducted their study on a sample of 161 Iranian startups based on accelerators and identified 39 variables affecting startup success. Interestingly, their two-staged stacking model yielded an impressive accuracy of 89%, indicating the potential of machine learning models in predicting startup success. Their study identified several key variables such as startup origin from accelerators, creativity and problem-solving abilities of founders, first-mover advantage, and the amount of seed investment, providing valuable insights for venture capitalists and decision-makers.

Building on this, (Thirupathi et al., 2021) adopted the XGBoost algorithm to predict the success of small businesses that received Small Business Innovation Research (SBIR) or Small Business Technology Transfer (STTR) awards. Their model achieved an accuracy of 84% and an AUC of 0.91, validating the efficacy of machine learning models in this domain. The study also highlighted the role of employees with entrepreneurial experience, arts, and/or STEM educational backgrounds in influencing business success. This research presents a novel approach to assessing the viability of small ventures and outlines key factors contributing to their success.

In a similar vein, (Zbikowski and Antosiuk, 2021) utilized machine learning algorithms to predict startup success, with the XGBoost algorithm achieving a precision score of 0.86. Their study identified a startup's location and industry as significant predictors of success, further expanding our understanding of the factors that influence startup success. This research underscores the potential of machine learning algorithms in offering valuable insights to investors and entrepreneurs.

Continuing this line of inquiry, (Abhinand and Poonam, 2022) machine learning techniques to identify factors impacting startup success in India. Their study achieved an accuracy of 80.1% with a stacked ensemble model, reinforcing the utility of machine learning models in predicting startup outcomes. Their research offers an insightful perspective on startup success in the Indian context, highlighting the global applicability of machine learning techniques.

(Srinivasan and P, 2020) took a different approach by focusing on the success of crowdfunding campaigns. They used an ensemble deep-learning model to achieve an impressive accuracy of 93%. Their study reveals the potential of combining textual and numeric features in predicting campaign success, opening a new avenue of research in the realm of crowdfunding and entrepreneurship.

On the other hand, (Pasayat et al., 2020) proposed a framework based on an evolutionary algorithm to identify crucial features related to startup success. Their innovative approach achieved an exceptional accuracy of about 92.3% when trained with popular machine learning classification frameworks. This study underscores the importance of feature selection and introduces a novel approach to predicting startup success, paving the way for future research in this area.

(Arroyo et al., 2019) examined how machine learning can improve venture capital investment decision-making. Using a dataset of over 120,000 early-stage companies from Crunchbase, the study aimed to predict possible outcomes over a 3-year time window, such as a funding round or closure of the company. The authors used several machines learning algorithms, including logistic regression, decision trees, random forests, gradient boosting, and neural networks, with the gradient boosting classifier achieving the highest F1-score of 0.63. The approach of predicting multiple outcomes instead of just two provides VC investors with more information to set up a lower risk portfolio with potentially higher returns. The study concludes that machine learning can support venture investors in their decision-making process to find opportunities and better assess potential investment risks.

Finally, (Ross et al., 2020) introduced a machine learning model called CapitalVX that predicts startup outcomes using a large dataset from Crunchbase and the USPTO. Achieving an out-of-sample accuracy of 88%, their model demonstrates the practical benefits of using machine learning to screen potential investments. This research shows how machine learning can optimize the investment process, freeing up time for mentoring and monitoring investments, thereby enhancing the efficiency and effectiveness of venture capital and private equity firms2.

In summary, these studies collectively demonstrate the power and potential of machine learning algorithms in predicting startup success. They elucidate the significant role of feature selection, highlight the key factors that influence startup success, and illustrate the practical implications of these predictive models. This body of research provides an invaluable resource for entrepreneurs, investors, and policymakers, offering data-driven insights to inform their decision-making and strategy development processes in the dynamic and complex world of startups.

# 4. Aim & Objectives

The aim of this thesis is to compare and evaluate the effectiveness of various machine learning models and techniques in predicting startup success. By exploring the factors that contribute to the success or failure of startups, this research aims to provide valuable insights for investors, entrepreneurs, and policymakers in making informed decisions about supporting and investing in startups. The study will build on existing research and contribute to the development of a more comprehensive model for accurately predicting startup outcomes.

**Objective**

To review the existing literature on startup success prediction and identify the key factors influencing it for USA.

To create a comprehensive dataset of startups, incorporating relevant features and success indicators.

To develop, train, and test various machine learning models for startup success prediction, such as logistic regression, support vector machines, decision trees, random forests, and deep learning techniques.

To conduct a comparative analysis of the performance of different machine learning models and techniques in predicting startup success.

To identify the most suitable machine learning models and techniques for accurately predict the success of startups.

# 5. Significance of Study

The significance of this study lies in its potential to contribute to the existing body of knowledge on startup success prediction. By comparing and evaluating various machine learning models and techniques, this research can provide valuable insights for investors, entrepreneurs, and policymakers in making informed decisions about supporting and investing in startups. The accurate prediction of startup outcomes can inform investment decisions and contribute to the growth of innovative businesses that can drive economic development. Additionally, this study can also contribute to the development of more comprehensive models that can effectively predict startup success or failure, which is crucial in the current business landscape.

# 6. Scope of Study

## In scope:

This thesis will focus on exploring the dynamics of startup success prediction using machine learning models and techniques. The study will specifically compare and evaluate the performance of various machine learning algorithms for predicting startup outcomes.

## Out of scope:

This study does not aim to provide an exhaustive list of factors that contribute to startup success or failure. It will also not cover the implementation of the proposed models in real-world scenarios.We are not taking online crunch base data if we needed, we could take that.

## Reason for defining the scope:

Defining the scope of the study will help ensure that the research remains focused and achievable within the given timeframe. By limiting the scope to the comparison of machine learning models and techniques for predicting startup success, the study can provide a comprehensive evaluation of these models' performance and inform investors, entrepreneurs, and policymakers about the most effective approaches for predicting startup outcomes.

# 7. Research Methodology

## Business Understanding

The understanding of business dynamics plays a pivotal role in the framework of this thesis. The investigation of factors contributing to business success and failure is crucial for the development of predictive models that can effectively anticipate future outcomes. Building upon existing research, it becomes evident that businesses undergo unique patterns of failure, necessitating a comprehensive analysis of failed firms to identify the key factors associated with their demise. The literature highlights the significance of both financial and non-financial variables in predicting business outcomes.

In the realm of predictive modelling, various machine learning algorithms have gained prominence due to their superior performance compared to traditional statistical models. This research explores the utilization of advanced machine learning algorithms such as ADYSN, XGBoost, logistic regression, decision trees, and random forests. These algorithms offer a flexible framework for analysing complex and diverse datasets, enabling the exploration of non-linear relationships, and alleviating the constraints imposed by conventional statistical models.

By integrating these cutting-edge machine learning techniques with a profound understanding of business dynamics, this research aims to provide a comprehensive comprehension of the multifaceted factors that shape business success and failure. It aspires to uncover valuable insights that can guide both practitioners and researchers, illuminating the path to success in the dynamic and ever-evolving business landscape.

## Metadata

In the domain of startup research, a common approach has been for researchers to formulate their own surveys and conduct interviews with startup stakeholders. This process generates direct data from both successful and struggling companies. However, such an approach has its limitations, mainly the constraint on the size of the dataset, which is often relatively small due to the time-consuming nature of the data collection process.

The primary objective of this research endeavour is to construct a reliable computational model that effectively predict success of a startup.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| 1. permalink | The unique identifier for the company, often in the format of a URL slug. |
| 2. name | The official name of the company. |
| 3. homepage\_url | The company's homepage URL. |
| 4. category\_list | The categories or industries the company operates in. |
| 5.funding\_total\_usd | The total amount of funding the company has received, in USD. |
| 6. status | The current operational status of the company (e.g., operating, acquired, closed). |
| 7. country\_code | The ISO country code of the company's location. |
| 8. state\_code | The state code of the company's location. |
| 9. region | The broader region where the company is located. |
| 10. city | The city where the company is located. |

**Data Processing:**

**7.4.3 Data Transformation**

Data Source

CrunchBase

Kaggle

Data Processing

EDA

Train Test Split

66367 Rows

14 Columns

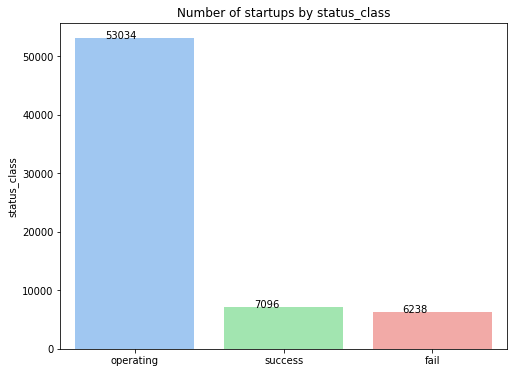
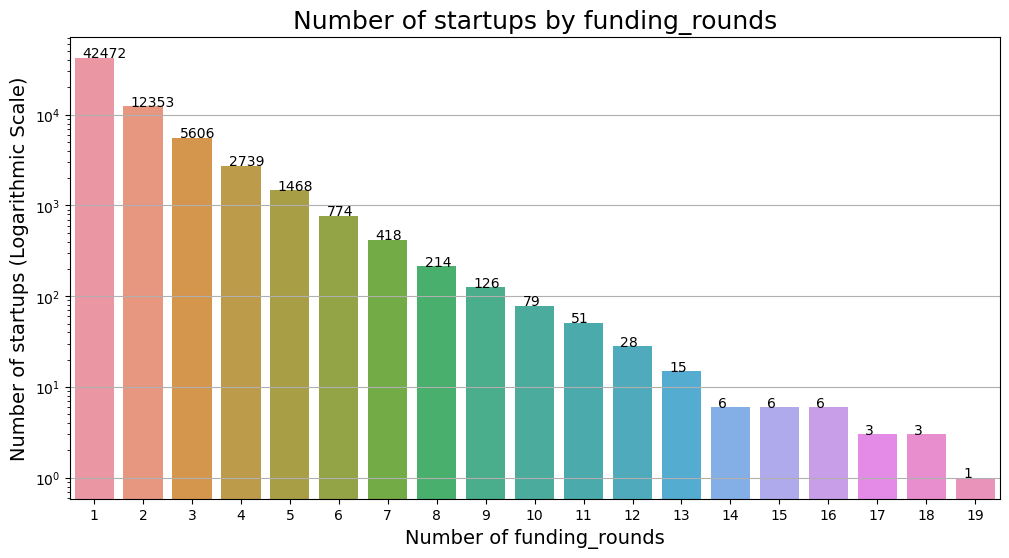
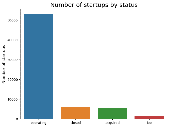
Model Training

Model Evaluation

Final Model

139 Countries

No of Funding rounds



Unique market column creation, irrelevant columns dropped, funding deferrals and operational years calculated, and average funding per round assessed.

filtering US data, handling missing values, and encoding variables.

The train-test split is performed with stratification to ensure the distribution of the target variable is preserved in both train and test sets.

Metrics and a logistic model are imported. Classification functions are defined and applied to multiple models.

XGBoost

Logistic Regression

Decision Tree

Random forest

Models are evaluated using Classification Report, ROC-AUC, Class Prediction Error, Cohen Kappa Score, and F1 Score

XGBoost

**7.4.4 Data Splitting**

**7.4.5 Feature Engineering**

**7.4.6 Data preparation for deep learning models**

**7.5 Model explanation**

# 8. Required Resources

Hardware and software requirements are as follow:

## Hardware requirements:

A computer with high processing power and a large storage capacity (at least 16GB of RAM, quad-core processor, and dedicated graphics card with at least 4GB of VRAM)500GB of storage space

## Software requirements:

* Utilize Python language (version 3.8+) for executing machine learning and data analytics.
* Incorporate TensorFlow (version 2.4+), a public library focused on deep learning and machine learning activities.
* Adopt Scikit-learn (version 1.0.3+), a Python-based toolkit for machine learning tasks.
* Implement Statsmodel (version 0.14.0+), a Python tool that enhances statistical calculations.
* Employ Matplotlib (version 3.7.1+), a visualization tool within Python.
* Integrate Seaborn (version 0.11.2+), a Python-centric data visualization toolkit.
* Use Pandas (version 2.0.2+), a Python tool designed for data processing.
* Incorporate Numpy (version 1.24.2+), a Python-based computational library.
* Utilize Jupyter Notebook (version 6.1+), an interactive platform compatible with Python.
* Maintain a system like Git (version 2.29+) to oversee and coordinate changes to both code and data over the course of the study.

It's possible that additional software tools and libraries may be required, depending on the specific needs of the research and the complexity of the models being developed. Overall, the required resources for this research proposal will include data, software, hardware, and expertise, and will enable the research to identify the potential benefits and challenges of using machine learning techniques in startup success prediction.

.

# 9. Research Plan



# References

Abhinand, G. and Poonam, B., (2022) An Efficient Stacking Ensemble Technique for Success Prediction of Indian Ventures. *2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI)*, 01, pp.1–6.

Arroyo, J., Corea, F., Jiménez-Díaz, G. and Recio-García, J.A., (2019) Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. *IEEE Access*, 7, pp.124233–124243.

Pasayat, A.K., Bhowmick, B. and Roy, R., (2020) Factors Responsible for the Success of a Start-up: A Meta-Analytic Approach. *IEEE Transactions on Engineering Management*, PP, pp.1–11.

Ross, G., Sciro, D., Das, S.R. and Raza, H., (2020) CapitalVX: A Machine Learning Model for Startup Selection and Exit Prediction. *ERPN: Entrepreneurs (Finance) (Topic)*, null, p.null.

Sadatrasoul, S., Ebadati, O. and Saedi, R., (2020) A Hybrid Business Success Versus Failure Classification Prediction Model: A Case of Iranian Accelerated Start-ups. *Journal of AI and Data Mining*, 8, pp.279–287.

Srinivasan, A. and P, A., (2020) An Ensemble Deep Learning Approach to Explore the Impact of Enticement, Engagement and Experience in Reward Based Crowdfunding. *ERPN: Entrepreneurs (Finance) (Topic)*, null, p.null.

Thirupathi, A.N., Alhanai, T. and Ghassemi, M.M., (2021) A machine learning approach to detect early signs of startup success. In: *Proceedings of the Second ACM International Conference on AI in Finance*. [online] ICAIF’21: 2nd ACM International Conference on AI in Finance. Virtual Event: ACM, pp.1–8. Available at: https://dl.acm.org/doi/10.1145/3490354.3494374 [Accessed 9 Apr. 2023].

Zbikowski, K. and Antosiuk, P., (2021) A machine learning, bias-free approach for predicting business success using Crunchbase data. *Inf. Process. Manag.*, 58, p.102555.

Top of Form