UNRAVELLING THE DYNAMICS OF STARTUP SUCCESS PREDICTION: A THESIS ON THE COMPARATIVE STUDY OF MACHINE LEARNING MODELS AND TECHNIQUES

ANKUR NAPA

Final Thesis Report

JUNE 2023

DEDICATION

This dissertation is a testament to my family's unwavering love and support, which has been instrumental to my academic journey. Their sacrifices made this research possible. I'm deeply grateful to my mentor, Shah Mohammad Azam, whose wisdom, guidance, and consistent encouragement profoundly shaped this research. This work reflects the steadfast support from my family, valuable mentorship from Shah Mohammad Azam, and the academic excellence at LJMU. I hope this research not only enhances existing knowledge but also inspires future researchers.

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 endeavour 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 centred 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

LIST OF ABBREVIATION

ADASYN	Adaptive synthetic oversampling technique
AUC	Area under the curve
FN	False negative
KNN	kth-nearest neighbor
SMOTE	Synthetic minority oversampling technique
TP	True positive
TPR	True positive rate
USD	US Dollars
XGB	Extreme gradient boosting

List of Figures

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 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.

1.2 Problem statement and Related work

1.2.1 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 endeavour.

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.

1.3 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.

1.4 Scope of Study

1.4.1 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.

1.4.2 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.

1.4.3 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.

1.5 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.

1.6 Structure of Study

This thesis focuses on the application of machine learning models, including logistic regression, decision tree, random forest, and XGBoost, to predict startup success using data from Crunchbase, and addresses class imbalance with ADASYN oversampling. Despite challenges in data collection, the models achieved high accuracy, and XGBoost emerged as the best performer. Enhancements, such as including richer data, redefining success measures, and harnessing real-time data, could further improve these predictive models.

CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

Defining startups presents a challenge due to the variation in criteria across different studies. However, common characteristics often include elements of novelty, activity, and independence. Nonetheless, the distinction between startups, spin-offs, and startups founded by larger corporations remains difficult due to limitations in available data. Despite this challenge, researchers have recognized the significance of integrating human expertise with machine learning models in predicting startup success. Several studies have focused on applying machine learning algorithms to forecast various aspects of startup performance and funding events, yielding valuable insights into the potential of these models and techniques. Ensemble techniques, which combine multiple models, have gained attention in the field of startup success prediction due to their ability to achieve higher accuracy rates compared to standalone models. Moreover, researchers have explored the influence of extrinsic factors such as competition and investor network positioning on startup fundraising success. Deep learning methodologies have emerged as promising tools for evaluating startups, showing potential in surpassing traditional statistical methods. However, the absence of a universally applicable approach to deep learning-based startup evaluation remains a challenge, as variations in data sources, processing techniques, and evaluation metrics hinder standardization. Addressing the challenges of data collection, measurement, and availability, innovative approaches utilizing publicly available web information have shown promise in predicting startup funding events. Nonetheless, further research is needed to overcome data limitations and enhance the accuracy of predictions. By synthesizing the current research landscape on startup success prediction, particularly within the realm of machine learning, this literature review contributes to our understanding of the dynamics and potential of machine learning models and techniques. It serves as a foundation for future exploration and comparison of various machine learning models and techniques in the context of predicting startup success.

(Pasayat et al., 2020) explores the advantages of using machine learning methods over conventional statistics to accurately predict a firm's financial performance. The authors highlight that machine learning algorithms can identify relevant variables in a dataset and integrate multiple data types from various sources to improve the accuracy of predictions. The approach used in the paper involves data-driven analysis and cross-validation, resulting in higher accuracy. However, the specific dataset used, list of algorithms, best-performing algorithm, and approach are not mentioned in the given information.

(Varma, 2021) explores the use of machine learning to predict the success of start-up companies. The authors suggest that while machine learning is a useful tool, it should be used in conjunction with human expertise. The paper uses data from Crunchbase, TechCrunch, Mattermark, and Dealroom, which was split into a training set and a testing set. Several machine learning algorithms were discussed, including logistic regression, decision trees, random forests, support vector machines, and neural networks. The random forest classifier was found to be the best-performing algorithm, achieving an accuracy of 85.7%. The paper highlights the importance of feature selection and identifies funding, industry, and location as key factors for predicting start-up success.

(Antretter et al., 2018) explores the use of digital traces to predict early-stage startup survival. They analysed the digital footprints of 542 entrepreneurs and used a text mining approach to predict 5-year survival rates with an accuracy of up to 91%. The paper also provides a taxonomy of important digital traces and benchmarks their approach against actual investments made by 339 business angels. The dataset used in the study is not explicitly mentioned, and the best performing algorithm is not specified.

(Bangdiwala et al., 2022) uses historical data available on startups to build and compare five machine learning models to predict if a startup would get acquired or not. The models used are Decision Trees, Random Forest, Gradient Boost, Logistic Regression, and MLP Neural networks, trained on a dataset obtained from CrunchBase. The dataset includes features such as valuations, funding rounds, and investments. The paper concludes that the models were able to achieve an accuracy of around 92% in predicting if a startup would get acquired or not. Additionally, the paper uses various columns from the dataset, including funding and investment details, founding individuals, and information about acquisitions and mergers, to provide insights into the latest trends and news of the startup industry.

(Li, 2020) analyses the success of startup companies using machine learning algorithms and a dataset containing information on 22,000 startups and their features. The authors use Random Forest and Support Vector Machine algorithms to classify startups and identify important features for success. The best-performing algorithm is Random Forest, achieving an 85.5% accuracy rate. Important features for success include funding rounds, number of employees, and industry sector. The paper concludes that machine learning methods can improve the efficiency of analysis and prediction for investors and venture capital companies.

(Ramalakshmi and Kamidi, 2018) introduces a machine learning model that utilizes key attributes at different stages of startup functioning to predict their funding range. Real-time data from the CrunchBase dataset between 2015 and 2017 is used, and classification and regression algorithms are applied to build the model. Compared to existing models, the proposed model demonstrates higher accuracy, providing valuable insights for investors and entrepreneurs when making funding decisions. Additionally, a technique is introduced to improve the model's accuracy by creating new columns from existing ones. The methodology involves data preprocessing, attribute identification, and training predictive models using machine learning algorithms and real-time data. Performance evaluation measures, including accuracy, precision, recall sensitivity, specificity, and area under the ROC curve, are employed to assess the model's effectiveness.

(Zhang et al., 2021) used Scalable Heterogeneous Graph Markov Neural Network (SHGMNN) is a novel system designed to pinpoint promising early-stage startups. It employs the Crunchbase dataset to gather pertinent details, converting them into graph-structured data. This approach entails the creation of various meta paths to encapsulate diverse semantics across the heterogeneous information network (HIN), which are then amalgamated into a comprehensive graph structure. Utilizing Graph Neural Network (GNN), the system diffuses this gathered information across the consolidated graph, while also employing Maximum A Posteriori (MAP) inference on Hinge-Loss Markov Random Fields to ensure label consistency. This results in a GNN with a lean linear diffusion design that carries out graph propagation across vast web-scale heterogeneous information networks. The practical application and experimental data on real-world datasets affirm the superiority of the SHGMNN approach.

(Böhm et al., 2017) conducted research on predicting the success of Indian startups using various machine learning approaches, including Random Forest, Naive Bayes, Decision Tree Classifier, and K-Nearest Neighbour. The data was collected from different sources, such as Venture Intelligence and Startup Talky, and was cleaned and pre-processed for visualization and algorithmic application by indexing it according to industry and headquarters location. Classifiers were used to compare metrics, and to improve accuracy, the authors proposed a stacked ensemble model. The ensemble model comprised standalone models, including Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbour, and Bernoulli Naïve Bayes Classifier, which were trained to produce intermediate predictions. These intermediate predictions were then fed into a Logistic Regression model, which was used to learn from them. The authors compared the performance of various models and found that the ensemble methods had the highest accuracy of 94.1%, with an AUC score of 92.22%. The reference for this paper is Ünal, C., G, A., & B, P. N. (2022). An Efficient Stacking Ensemble Technique for Success Prediction of Indian Ventures. 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI).

(Kaiser and Kuhn, 2020) delves into the utilization of publicly accessible, web-extracted data to forecast the performance trajectories of nascent enterprises over a half-decade span. They employ a unique blend of textual and non-textual information, such as the names and addresses of Danish firms, coupled with fundamental accounting data and characteristics of the founders and the startups, to anticipate various performance outcomes. These outcomes include instances of involuntary exit, surpassing average employment growth, achieving over 20% return on assets, filing new patent applications, and enrolling in an innovation subsidy scheme. The predictive models, overall, demonstrate either high or exceptionally high accuracy, save for predicting high returns on assets, where the predictive efficacy remains relatively weak. The robust precision of these forecasts for outcomes such as survival, employment growth, patent filings, and participation in innovation support schemes, underscores the potential for algorithmic scoring models as an instrumental complement to decisions pertaining to funding and innovation support. The data leveraged in the study was availed through the project "Investments, Incentives and the Impact of Danish Research", underwritten by the Novo Nordisk Foundation. A variety of algorithms, such as logistic regression, random forests, and gradient boosting were used, with gradient boosting outperforming the rest in predicting startup performance outcomes. In terms of data preprocessing, the authors utilized an array of techniques like text sanitization, stemming, and elimination of stop-words. They also devised indicators to track the business history linked to each address and to identify addresses that were shared with other firms.

(Böhm et al., 2017) the novel concept of 'Business Model DNA', designed to encapsulate the distinctive attributes of business models. This innovative framework enables the scrutiny of business models to detect clusters that outperform others and to anticipate the potential trajectories of specific business models. This analysis leverages data from 181 startups based in the USA and Germany, using data mining methodologies such as cluster analysis and Support Vector Machines to categorize various business models based on their performance. The authors unearthed 12 unique clusters of business models, each exhibiting divergent growth expectations and survival probabilities. The introduced model demonstrates a capability to predict a venture's survival with a commendable accuracy of 83.6%. While the paper underscores the considerable value of the proposed model for startup classification and success prediction, it also underscores the necessity for more data to conclusively validate, augment, and fine-tune the model. Despite the absence of any mention of specific techniques employed for data preprocessing, the study utilized data mining techniques, such as cluster analysis and Support Vector Machines, to classify different business models in terms of their performance. The paper, however, does not explicitly identify the best performing algorithm used in the study.

(Gastaud et al., 2019) investigates the determinants that propel the success of startup fundraising, focusing on two extrinsic characteristics: competition and the network position of investors. Utilizing data gathered from the Crunchbase database, the study applies the Word2Vec algorithm to gauge the degree of competition and employs Graph Neural Networks (GNN) to scrutinize network characteristics. The findings underscore that competition significantly influences fundraising during the early stages, whereas network features gain prominence during the growth stage. The authors conclude that the triumph of startup fundraising is governed by a blend of intrinsic and extrinsic factors. Although intrinsic elements like the characteristics of the founders are vital, extrinsic factors, namely competition and the network position of investors, also wield substantial influence. Consequently, the authors caution against over-reliance on global models that average these factors, as they may distort the true picture of startup success in fundraising. In terms of methodology, the authors utilized the Word2Vec algorithm and Graph Neural Networks (GNN) in their analysis but did not specify the best performing algorithm. For data preprocessing, the authors employed the Yeo-Johnson's transformation to standardize the data and enhance the normality of the features, alongside word embeddings and cosine similarity to measure competition.

(Cao et al., 2022) delves into the exploration of deep learning (DL) for assessing the likelihood of startup success. Through a comprehensive literature review and synthesis of DL-based methodologies, spanning the complete DL lifecycle, the authors aim to acquire an exhaustive understanding of the tactics for startup evaluation employing DL and to extract insightful and practical knowledge for practitioners. The study concludes that DL-oriented strategies hold the potential to surpass traditional statistical methods in startup evaluation. Yet, it highlights the absence of a universally applicable approach to DL-based startup evaluation, given the uniqueness of each work in aspects like data sources, modalities, processing, feature engineering, success criteria for startups, evaluation metrics, and so forth. While the paper does not reference any specific data source or list of algorithms used in the study, it underscores the significance of gathering suitable data for model input and making informed choices about the source, type, modality, and size of the data. Moreover, it elucidates the challenges of reaching a definitive conclusion about the superior DL model due to the unique attributes of each work. In terms of data preprocessing for DL-based startup evaluation, the authors discuss several special techniques such as normalization, augmentation, debiasing, balancing, and densifying, which can enhance the performance of DL models.

(Garkavenko et al., 2022) explores the potential of predicting a startup's ability to secure investments using freely accessible public data. The authors develop an innovative approach that relies exclusively on readily available sources of information, such as a startup's website, its social media engagement, and its general web presence, to anticipate its funding events. Remarkably, this approach manages to produce results on par with those that also employ structured data from private databases. The authors conclude that their proposed methodology can proficiently predict startup funding events using only freely available web information, which could potentially guide investors in making well-informed decisions. The data sources employed in the study encompass freely available web information like a startup's website, its social media activity, and its overall web presence. Regarding algorithms, the study explores Logistic Regression and CatBoost to forecast new funding events, with CatBoost outperforming the former. Despite the absence of a specific mention of special techniques used for data preprocessing, the study significantly contributes to the domain of startup investment prediction using publicly available web data.

2.2 Startup prediction using ML Models

2.2.1 Random Forests

Random Forests is a machine learning algorithm that consists of numerous decision trees operating as an ensemble. Each individual tree in the random forest algorithm produces a class prediction, which is considered a "vote", and the class with the most votes becomes the model's prediction. This algorithm, due to its capability of handling a large volume of data and providing an appreciable accuracy, has been extensively used in numerous studies for startup performance prediction.

(Varma, 2021) utilized Random Forest in their research and found it to be the most effective algorithm, boasting an impressive accuracy of 85.7%. This study shows the potential of Random Forests in identifying promising startups by accurately classifying them based on various features. Similarly,(Li, 2020) also reported that the Random Forest algorithm was the best-performing algorithm in their study, closely mirroring Varma's results with an accuracy rate of 85.5%.

Moreover, (Böhm et al., 2017) incorporated Random Forest along with other algorithms for their study. Interestingly, they observed that the ensemble methods, which combine the outputs of several base models, had the highest accuracy. This suggests that combining the predictions of multiple models, including Random Forests, can potentially enhance the predictive performance, especially when dealing with complex datasets. In the same study, they also used Random Forest in conjunction with Support Vector Machines and other data mining methodologies for the tasks of classification and performance prediction.

However, it is essential to note that while Random Forest has shown robust performance in many scenarios, there are instances where other algorithms have outperformed it. For instance, (Kaiser and Kuhn, 2020) reported that while they did utilize Random Forest in their research, gradient boosting was found to outperform all algorithms when predicting startup performance outcomes.

These studies provide evidence of the efficacy of the Random Forest algorithm in startup success prediction and its ability to handle diverse and complex data. However, they also suggest that the choice of the algorithm should be context-dependent and that other algorithms such as gradient boosting might be more effective in certain scenarios.

2.2.2 Decision Trees

Decision Trees are a popular machine learning algorithm that employs a tree-like model to make decisions or predictions based on a set of input features. In this algorithm, the dataset is recursively split into subsets based on the values of specific attributes, resulting in a hierarchical structure of decision nodes and leaf nodes. Each decision node represents a test on an attribute, while each leaf node represents a class label or a prediction.

In the context of startup prediction and performance analysis, Decision Trees have been utilized in several studies to assess their effectiveness. (Varma, 2021) incorporated Decision Trees as one of the algorithms in their research. Although specific details regarding the use of Decision Trees were not provided, their inclusion suggests that Decision Trees were deemed relevant and potentially effective in predicting startup outcomes.

Similarly, (Bangdiwala et al., 2022) employed Decision Trees as one of the models in their study. While no further information was given regarding the specific implementation or performance of Decision Trees, their inclusion indicates the recognition of Decision Trees as a viable approach for startup prediction tasks.

Furthermore, (Böhm et al., 2017) utilized the Decision Tree Classifier in their research. The Decision Tree Classifier was employed as an individual model as well as part of an ensemble model. The ensemble model incorporated multiple base models, including Decision Trees, to improve predictive accuracy and performance in the classification and prediction of startup outcomes.

These studies collectively highlight the utilization of Decision Trees in startup prediction and performance analysis. Although the specific details and outcomes of using Decision Trees vary across the studies, their inclusion underscores their potential as a valuable algorithm for making decisions and predictions in the context of startups.

2.2.3 Logistic Regression

Logistic Regression is a widely used statistical modelling technique that is employed to predict binary or categorical outcomes based on a set of input variables. It is particularly suitable for analysing the relationship between a dependent variable and one or more independent variables by estimating the probabilities of different outcomes.

The application of Logistic Regression in the prediction of startup outcomes has been explored in several studies. (Varma, 2021) incorporated Logistic Regression as one of the algorithms in their research. Although specific details regarding the implementation and performance of Logistic Regression were not provided, its inclusion suggests the recognition of its relevance and potential effectiveness in startup prediction tasks.

Similarly, (Bangdiwala et al., 2022) employed Logistic Regression as one of the models in their study. Although no further details were provided regarding the specific implementation or performance of Logistic Regression, its inclusion indicates the acknowledgment of its usefulness as a modelling technique for startup prediction.

Furthermore, (Böhm et al., 2017) utilized a Logistic Regression model within an ensemble model. The ensemble model combined the predictions of multiple base models, including Logistic Regression, to enhance the learning process and improve predictive performance.

However, it is important to note that in some instances, other algorithms have outperformed Logistic Regression. (Kaiser and Kuhn, 2020) reported that while they did employ Logistic

Regression in their research, gradient boosting emerged as the superior algorithm in predicting startup performance outcomes.

Additionally, (Garkavenko et al., 2022) used Logistic Regression in their study, but CatBoost, another algorithm, outperformed it in terms of predictive performance.

These studies collectively highlight the application of Logistic Regression in the context of startup prediction. While its specific implementation and performance outcomes may vary across studies, the inclusion of Logistic Regression underscores its relevance and potential as a modelling technique for analysing and predicting startup outcomes.

2.2.4 Support Vector Machines (SVM)

Support Vector Machines (SVM) is a powerful machine learning algorithm commonly used for classification and regression tasks. SVM aims to find an optimal hyperplane that separates data points belonging to different classes with the largest margin. It is particularly effective when dealing with high-dimensional datasets and non-linear decision boundaries.

In the context of startup prediction and performance analysis, SVM has been employed in several studies, demonstrating its usefulness and versatility. (Varma, 2021) utilized SVM as one of the algorithms in their research. Although specific details regarding the implementation and performance of SVM were not provided, its inclusion suggests that SVM was considered a relevant algorithm for startup prediction tasks.

Similarly, (Li, 2020) reported the use of SVM in their study; however, Random Forest outperformed SVM in terms of predictive performance. This suggests that while SVM was employed, it did not yield the best results in their specific analysis.

Furthermore, (Böhm et al., 2017) utilized SVM in conjunction with Random Forest and other data mining methodologies for the classification and prediction of startup outcomes. The combination of SVM and Random Forest, along with other algorithms, showcased the potential of utilizing multiple models and techniques to enhance predictive accuracy and performance.

These studies collectively emphasize the application of SVM in startup prediction tasks. While its specific implementation and performance outcomes may vary across studies, the inclusion of SVM underscores its effectiveness and suitability for handling classification and performance prediction challenges in the context of startups.

2.2.5 Neural Networks

Neural Networks are a class of machine learning algorithms inspired by the structure and function of the human brain. They consist of interconnected layers of artificial neurons, where each neuron processes input data and passes it to the next layer, eventually producing an output. Neural Networks are particularly effective in capturing complex patterns and relationships in data, making them suitable for tasks such as image recognition, natural language processing, and prediction.

In the context of startup prediction and performance analysis, Neural Networks have been utilized in several studies, highlighting their ability to learn and extract meaningful features from data. Varma (2021) incorporated Neural Networks as one of the algorithms in their research. Although specific details regarding the implementation and performance of Neural Networks were not provided, their inclusion suggests the recognition of their relevance and potential effectiveness in startup prediction tasks.

Similarly, (Bangdiwala et al., 2022) employed MLP (Multi-Layer Perceptron) Neural Networks as one of the models in their study. MLP Neural Networks, characterized by multiple layers of interconnected neurons, are capable of learning complex relationships and patterns in data. The utilization of MLP Neural Networks indicates the acknowledgment of their suitability for modelling startup prediction tasks.

2.2.6 Gradient Boost

Gradient Boosting is a powerful ensemble learning technique that combines multiple weak predictive models, typically decision trees, to create a strong and accurate predictive model. It works by iteratively improving the predictions of each weak model, with each subsequent model attempting to correct the errors made by the previous models. Gradient Boosting has shown remarkable success in various prediction tasks, including startup performance prediction.

In the study conducted by (Bangdiwala et al., 2022) Gradient Boost was employed as one of the models in their research on startup prediction. The specific details of the implementation and performance of Gradient Boost were not provided, but its inclusion indicates its recognition as a suitable and potentially effective modelling technique.

Moreover, (Kaiser and Kuhn, 2020) reported that Gradient Boosting outperformed all other algorithms in predicting startup performance outcomes. This observation highlights the superiority of Gradient Boosting in capturing complex relationships and making accurate predictions in the context of startup analysis.

2.2.7 Naive Bayes

Naive Bayes is a probabilistic classification algorithm that applies Bayes' theorem with the assumption of independence between features. Despite its simplicity and the naive assumption, Naive Bayes has demonstrated its effectiveness in various classification tasks. It is particularly well-suited for text classification and spam filtering applications.

In the study conducted by (Böhm et al., 2017) Naive Bayes was one of the algorithms used for startup prediction. Although further details regarding the implementation and performance of Naive Bayes were not provided, its inclusion indicates the recognition of its potential effectiveness in capturing patterns and making predictions in the context of startups.

2.2.8 K-Nearest Neighbor

K-Nearest neighbour (KNN) is a non-parametric classification algorithm that classifies new data points based on the majority class of their neighbouring data points in the feature space. KNN is a simple yet effective algorithm for both classification and regression tasks. It does not require explicit training and can adapt to changing data distributions.

(Böhm et al., 2017) utilized K-Nearest neighbour as one of the algorithms in their startup prediction research. The specific details and performance outcomes of K-Nearest neighbour were not provided, but its inclusion suggests its recognition as a relevant algorithm for capturing similarities and making predictions in the context of startups.

2.2.9 Other Classification and Regression Algorithms

In the study conducted by (Ramalakshmi and Kamidi, 2018) other classification and regression algorithms were used for startup prediction. However, the specific algorithms employed were not explicitly mentioned. The utilization of a variety of classification and regression algorithms suggests.

2.3 Literature Review of proposed models and algorithm

2.3.1 ADASYN

(Haibo He et al., 2008) offers a valuable tool for improving startup success prediction. By addressing the challenges posed by imbalanced datasets, where the minority class (successful startups) is underrepresented, ADASYN enables the generation of synthetic data points for the more difficult-to-learn minority class examples. This rebalancing of the dataset reduces bias and enhances the performance of machine learning algorithms in predicting startup success. Integrating ADASYN into the prediction process involves preprocessing the dataset to create a more balanced representation of successful startups through synthetic sample generation. This approach enhances the accuracy of models and facilitates a better understanding of the factors driving startup success. Incorporating ADASYN into other prediction empowers researchers and practitioners to overcome data imbalance challenges.

2.3.2 Logistic Regression

(Ali and Jabeen, 2022) employed logistic regression analysis to uncover the underlying determinants influencing individuals' propensity for embarking on entrepreneurial ventures. Logistic regression, a statistical technique, was chosen to examine the association between a dependent variable, start-up intention, and various independent variables. The independent variables encompassed demographic characteristics, attitudes towards entrepreneurship, subjective norms regarding entrepreneurship, and perceived behavioural control in relation to entrepreneurship. Through the logistic regression model, the researchers calculated the likelihood of start-up intention by considering these independent variables. The regression equation was expanded to incorporate the factors influencing the inclination towards starting a business.

2.3.3 Decision Trees

(Wang et al., 2020) aimed to predict the fundraising outcomes of crowdfunding projects by utilizing both deep learning and commonly used machine learning algorithms. The study findings revealed that the deep learning model exhibited superior predictive performance, followed by the decision tree model. This investigation highlights the significant advantages of deep learning across various evaluation criteria, demonstrating its potential in forecasting crowdfunding project financing. The research also combined machine learning techniques with Internet finance, providing valuable insights and practical implications for future studies. To achieve their research objectives, the authors incorporated the decision tree model, a widely employed machine learning algorithm, for predicting crowdfunding fundraising outcomes. The implementation of the decision tree model was carried out using the scikit-learn library, and parameter tuning was performed using the grid search method. Cross-validation techniques were utilized to mitigate the risk of overfitting, and optimal parameters for the decision tree model were determined through the grid search process.

2.3.4 Random Forest

(Abhinand and Poonam, 2022) aims to employ machine learning techniques to predict the success of Indian startups. The researchers gathered data by web scraping multiple websites, compiling a dataset featuring various features describing both unicorn and non-unicorn startups in India. They utilized several machines learning classifiers, including Naïve Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest neighbour, to forecast the success of startups.

Furthermore, they explored the efficacy of a stacked ensemble model incorporating all the aforementioned classifiers to enhance prediction accuracy. The models underwent evaluation using metrics such as Accuracy, F1-Scores, and AUC scores. Results revealed that the Random Forest Classifier exhibited the highest accuracy of 78.8% and an AUC score of 0.79. Additionally, the stacked ensemble model achieved an accuracy of 80.1%. The researchers applied the model to predict the success rate of startups featured in the reality show, Shark Tank India, during its first season. The study's findings suggest that the proposed model can effectively predict the success of Indian startups, serving as a valuable tool for investors in making informed investment decisions.

2.3.5 XGBoost

(Agarwal, 2023) proposes a machine learning-based solution to predict the success of a startup. The authors use ensemble methods such as LGM Classifier, XGBoost, and AdaBoost Classifier for training and SVM for classification instead of the traditional SoftMax function. They also use margin loss instead of a standard entropy-based algorithm for the loss function and SMOTE and Tomek's links for balancing the data as the dataset is imbalanced. The approach produces 90.2% accuracy, precision, and F1 score, significantly improving various existing models. Therefore, XGBoost can be used as one of the ensemble methods for training the model to predict the success of a startup.

2.4 Research Gap

Embarking on a quest to address the gaps in existing research, we propose a comprehensive approach to accurately predict startup success. This approach leverages XGBoost, Random Forest, and ADASYN, and is designed to be applicable globally. We aim to use ensemble learning methods to enhance prediction accuracy, with feature selection made efficient by XGBoost's innate feature importance metric. The interpretability of these models fosters a synergistic relationship with human expertise, while their capacity for multi-class classification allows for nuanced outcome predictions. Finally, the use of ADASYN overcomes the challenge of imbalanced datasets, ensuring robust and reliable predictions. In essence, we seek to provide a reliable, globally relevant, and nuanced prediction model for startup success using advanced machine learning methodologies.

2.5 Summary

In this section, an exploration of various machine learning techniques for forecasting the success of startups is presented. It covers the utilization of Random Forests, Decision Trees, Logistic Regression, Support Vector Machines (SVM), and Neural Networks. The efficacy of each technique is assessed based on its capacity to manage diverse and intricate data, as well as its precision in predicting startup outcomes.

Random Forests have demonstrated strong performance in numerous situations, with research indicating an impressive accuracy rate of approximately 85.7%. However, there have been instances where other techniques, such as gradient boosting, have surpassed it. While Decision Trees and Logistic Regression have proven to be effective, they have also been outperformed by other techniques in certain circumstances.

SVM has been used in a multitude of studies, showcasing its versatility and effectiveness, particularly when dealing with high-dimensional datasets and non-linear decision boundaries. Neural Networks, specifically Graph Neural Networks (GNN), have been utilized to examine network features and competition, which have a significant impact on fundraising during a startup's early stages.

This section also introduces the innovative concept of 'Business Model DNA', a framework designed to encapsulate the unique attributes of specific business models. This groundbreaking framework allows for the examination of business models to identify clusters that outperform others and predict the potential trajectories of specific business models. In conclusion, this section underscores the significance of selecting algorithms based on context and the potential of machine learning in predicting startup success. However, it also highlights the necessity for additional research to pinpoint the most effective algorithm for particular scenarios.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

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.

Table 1 Description of given data

3.2 Research Methodology

3.2.1 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.

3.2.2 Data Selection

This study embraces a distinct approach by harnessing the power of machine learning algorithms to analyse a significantly expansive dataset, thereby facilitating a more comprehensive and resilient prediction of startup success. To accomplish this, we sourced our data from an open-source dataset available on Kaggle, which in turn originates from crunchbase.com, an extensive repository of startup information. This open-source data set offers a wealth of diverse and comprehensive data, amplifying the breadth and dependability of our predictive analyses.

3.2.3 Data Pre-processing

The initial step involves narrowing down the data to only comprise startups whose **'country_code'** identifies them as based in the United States. This allows for a concentrated study on the U.S. startup environment.

Data points lacking information about the **'funding_rounds'** are omitted from the study. It's essential to have this information as it can provide key insights into the startup's financial journey and growth trajectory.

Regarding the '**status**' of startups, those flagged as **'closed'** are treated as unsuccessful cases, while all others are assumed to have succeeded. This provides a clear dichotomy to gauge success and failure.

The study excludes businesses that were established before the year 2009, according to the **'founded_at'** field. This is to focus on newer businesses that still fall under the category of startups.

Any startup lacking vital details such as **'founded_at'**, 'name', or **'homepage_url'** are not considered in the study. The absence of these details raises concerns about the legitimacy of the business, potentially being ghost firms.

Redundant entries within the dataset are pruned to avoid any overemphasis or repetition of particular data points.

Startups without a specified 'region' are removed from the dataset. Regional data is crucial as it offers insights into the locational factors affecting the startup's performance.

The data is rigorously cleaned to remove any statistical anomalies or outliers that could possibly distort the results of the study.

Features that exhibit zero or near-zero variance are eliminated from the dataset. These features add little value to the predictive model due to their lack of variability.

For the **'funding_total_usd'** column, any instances of "-" are replaced with a zero USD value. This step ensures numerical consistency, thereby facilitating accurate computational analysis.

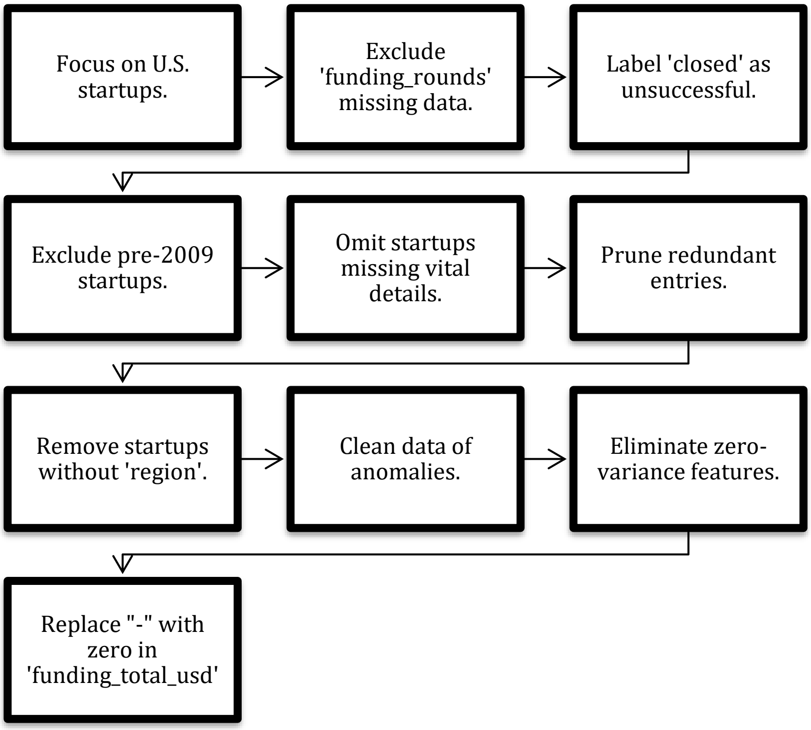

Focus on U.S. startups.
Exclude 'funding_rounds' missing data.
Label 'closed' as unsuccessful.
Exclude pre-2009 startups.
Omit startups missing vital details.
Prune redundant entries.
Remove startups without 'region'.
Clean data of anomalies.
Eliminate zero-variance features.
Replace "-" with zero in 'funding_total_usd'

Figure 1 : Data Preprocessing steps

3.2.4 Data Transformation

3.2.4.1 Variable Selection and Enhancement

Market Column: To enhance our analysis, we devised a unique approach by creating a dedicated market column. We carefully considered the category list, giving particular emphasis to the first category mentioned, as it often provided a significant indication of the startup's primary market focus. For cases where multiple categories were mentioned, we employed a strategic method to consolidate and synthesize this information, resulting in a more refined and informative market column.

Variable name	Description
funding_total_usd	This represents the total amount of money a startup has received from investors, typically expressed in US dollars.
status	This field indicates the current operating status of the startup. Common statuses include "operating", "acquired", "closed", etc.
country_code	This is a code that represents the country where the startup is headquartered.
state_code	Similar to the country code, this is a code that represents the state (within a country) where the startup is located.
region	This refers to the geographical region where the startup is located. It is often more specific than country and can refer to an area within a state or a metropolitan area.
funding_rounds	This field represents the number of distinct rounds of funding the startup has gone through, such as Seed, Series A, Series B, etc.
founded_at	This is the date when the startup was officially established or founded.
first_funding_at	This is the date when the startup received its first round of funding from investors.
last_funding_at	This represents the date when the startup received its most recent round of funding.
founded_Year	This field typically represents the year the company was founded.
final_status	This field could potentially represent the final operational status of the startup (for instance, whether it is still operating, has been acquired, or has shut down). The exact definition might vary depending on the specific database.
market	This refers to the market or industry in which the startup operates, such as technology, healthcare, finance, etc.
funding_diff	This could potentially be the difference between two funding rounds or between the first and last funding amount. The exact definition may vary depending on the specific database.

Table 2 Final processed metadata with new columns

Permanent Page, Home Page Links, Category List: During the data preprocessing phase, we made deliberate decisions to drop certain columns from the dataset. This included the permanent page, home page links, and category list columns. These columns were deemed less relevant for our specific research objectives, as they did not directly contribute to our analysis of startup success prediction. By focusing on the core variables of interest, we aimed to streamline our dataset and enhance the overall clarity and coherence of our findings.

Funding Deferrals Value: A vital aspect of our analysis involved assessing the duration between the first and last funding rounds, denoted as the funding deferrals value. By calculating the time span in which startups secured subsequent rounds of funding, we gained insights into their ability to attract additional financial support. This metric allowed us to gauge the frequency and timing of funding milestones, shedding light on the sustainability and growth potential of the startups under examination.

Years of Operating: To establish a clear understanding of each startup's operational tenure, we computed a variable called "years of operating." This involved subtracting the year of formation from the current year, providing us with a robust measure of the startup's duration in the market. By quantifying the number of years, a startup has been actively operating, we obtained valuable insights into their level of experience, market presence, and adaptability.

Funding Amount per Round: To assess the average funding amount received per funding round, we divided the total funding amount by the total number of funding rounds. This measure offered insights into the financial support each startup received during each funding stage, providing a perspective on the funding landscape and investment patterns.

These unique approaches and preprocessing steps within our analysis framework contribute to a comprehensive and nuanced examination of startup success prediction. By refining and transforming the data through innovative methods, we aim to uncover valuable insights that can inform strategic decision-making and foster a deeper understanding of the factors influencing startup performance and longevity.

3.2.4.2 One hot encoding

Market Column:

To transform the categorical variable "market" into a numerical format suitable for machine learning algorithms. One-hot encoding will be applied to the "market" column, creating separate binary columns for each unique market category. This encoding technique will enable the algorithms to effectively process and interpret the market data.

Country Code:

To convert the categorical variable "country code" into a numeric representation for modeling purposes. Similar to the market column, we will employ one-hot encoding on the "country code" column. This will generate a series of binary columns, each corresponding to a specific country code. By transforming the country code into a numerical format, the algorithms can better capture any potential relationships or patterns related to geographical factors.

Regions:

To transform the categorical variable "regions" into a format suitable for analysis and modeling.
We will utilize one-hot encoding on the "regions" column to create separate binary columns representing each unique region. This encoding technique will facilitate the inclusion of regional information in our models, allowing for potential insights into geographic variations and their impact on the target variable.

By applying one-hot encoding to the "market," "country code," and "regions" columns, we aim to convert these categorical variables into numerical representations that can be effectively utilized by machine learning algorithms. This transformation enables us to capture the underlying patterns and relationships within these variables, enhancing the predictive power of our models.

3.3 Class balancing

Class balancing is crucial in machine learning predictions to mitigate bias, improve model performance, and ensure fair treatment of all classes. It helps prevent the model from favoring the majority class, allows for accurate evaluation metrics, and enables effective generalization to unseen data and rare class instances.

3.3.1 ADASYN (Adaptive Synthetic Sampling)

It is an algorithm used for addressing class imbalance in machine learning. It dynamically generates synthetic samples for the minority class, emphasizing the harder-to-learn instances to achieve a more balanced representation, thereby improving model performance and addressing the challenges posed by imbalanced datasets.

3.4 Models

3.4.1 Logistic Regression

As per (Zou et al., 2019) Logistic Regression is a statistical model used for predicting the probability of binary outcomes, where the dependent variable can take only two values, such as 0 and 1. This method is extensively utilized in various sectors like healthcare, finance, and marketing, to predict outcomes like disease diagnosis, credit risk, and customer churn.

In its fundamental form, it uses a logistic function, also known as the sigmoid function, to model the binary dependent variable. This modeling is typically done in scenarios where the dependent variable is categorical, usually binary (e.g., yes/no, success/failure). The independent variables in this model can be continuous, categorical, or a mix of both.

Let's assume we have a dataset with 'n' observations, each having two features, X1 and X2. The goal is to predict a binary outcome Y (0 or 1). The model first computes a linear combination of the features, weighted by the learned coefficients from the training data:

$$\mathbf{z = \beta0 + \beta1X1 + \beta2X2}$$

Here, β0, β1, and β2 are the parameters of the model.

Then, it applies the sigmoid function to this output 'z' to yield the predicted probability 'p' of the positive class:

p = 1 / (1 + e^-z)

This resultant value 'p' is a probability that ranges between 0 and 1. A threshold value, often 0.5, is selected to make a binary prediction. If 'p' is above this threshold, the model predicts the positive class (Y=1). If it falls below, the model predicts the negative class (Y=0).

The model is trained using a method called maximum likelihood estimation, which identifies the regression coefficients that maximize the likelihood of the observed data. The model's performance is assessed using metrics such as accuracy, precision, recall, and the F1 score.

Logistic regression has the advantage of offering probabilities, making it interpretable. It is also relatively simple and fast, making it an excellent baseline for binary classification problems. However, it may struggle with datasets where the decision boundary is not linear, and it assumes that the features are independent, which might not always be the case in real-world scenarios.

3.4.2 Decision Tree

A decision tree, resembling a flowchart, is a hierarchical structure utilized primarily for classification tasks. Each internal node in this structure signifies a test on a particular attribute, every branch symbolizes the result of such a test, and each terminal node or leaf node contains a class label. The root node is located at the very top of the tree.

Building a decision tree doesn't involve a straightforward mathematical formula due to its conditional, tiered structure. Nonetheless, the formation of a decision tree can be articulated using concepts from information theory such as entropy and information gain, employed by algorithms like ID3 and C4.5, or Gini impurity, used by the CART (Classification and Regression Tree) algorithm.

Entropy is a concept used to quantify the disorder or impurity in a given set, defined as:

$$E(S) = - \sum pi * \log2(pi) \text{ for } i = 1 \text{ to } n$$

where `S` denotes the total sample space and `pi` represents the probability of selecting a sample of class `i` from `S`.

Information gain, on the other hand, measures the reduction in entropy. The attribute providing the highest information gain is selected for decision-making (splitting). It's defined as:

$$IG(S, A) = E(S) - \sum [|Sv| / |S| * E(Sv)] \text{ for } v \in \text{all values of attribute } A$$

where `A` symbolizes an attribute, `S` is the total sample space, `Sv` is the subset of `S` for which attribute `A` has value `v`, and `|Sv|` and `|S|` are the cardinalities of sets `Sv` and `S`, respectively.

For the CART algorithm, Gini impurity is used as a measure of impurity or purity. It's defined as:

$$Gini(S) = 1 - \sum (pi)\text{^}2 \text{ for } i = 1 \text{ to } n$$

where `S` is the total sample space and `pi` is the probability of selecting a sample of class `i` from `S`.

These metrics guide the decision on which attribute to split. The decision tree algorithm will repeat this procedure, creating divisions and advancing down the branches until it meets a predetermined stopping criterion. This could be when there are no more attributes to split on, all samples at a node belong to the same class, or a pre-set maximum depth is reached.

This decision tree concept is extensively discussed by (Charbuty and Abdulazeez, 2021) in his comprehensive review paper on decision tree algorithms applied in various machine learning domains. (Charbuty and Abdulazeez, 2021) discussed the types, advantages, and limitations of decision tree algorithms, with a focus on how these algorithms utilize information gain, also known as mutual information, as a metric for segmentation. The gain in information is defined based on the concept of entropy, as previously explained. He further delves into comparing different approaches to decision tree algorithms and discusses their uses with different types of datasets and associated findings. This analysis underscores the versatility and adaptability of decision tree algorithms in handling a range of classification tasks, despite their challenges.

3.4.3 Random Forest

(Zhu et al., 2019) random forest is a powerful supervised learning algorithm that is adept at performing a variety of tasks, including classification and regression. This algorithm operates by creating a multitude of decision trees during the training phase, with the final output being the mode of the classes for classification tasks or the mean prediction for regression tasks from individual trees.

Conceptually, a Random Forest comprises numerous unrelated decision trees. The Gini index, a statistical measure of inequality, serves as the splitting attribute selection metric for the decision trees. The number of levels in each branch of the tree depends on the parameter 'd' of the algorithm.

The algorithm constructs several decision trees, each using a random subset of the training data and a random subset of features. This randomness in data and feature selection for each tree helps to mitigate overfitting, thereby enhancing the accuracy of the model. After the decision trees are trained on these different subsets, the algorithm combines their predictions to make a final decision. This mechanism also allows the Random Forest algorithm to rank feature importance and simplify the learning process to optimize model computation.

Mathematically, Random Forest does not follow a simple formula like Logistic Regression or a Decision Tree due to its ensemble nature. However, the underlying operating principle can be represented mathematically.

In a Random Forest, there are 'n' decision trees. Each tree produces a classification outcome, denoted as Yj, where 'j' is the tree index. For a binary classification problem, each Yj can be 0 or 1.

The output of the Random Forest classifier, Y_RF, is the majority vote (for classification) or the average (for regression) of the outputs from individual decision trees. In the case of classification, this can be mathematically written as:

$$\mathbf{Y\_RF = mode(Y1, Y2, ..., Yn)}$$

For a regression problem, the output is the average of the outputs of individual trees:

$$\mathbf{Y\_RF = (Y1 + Y2 + ... + Yn) / n}$$

It's critical to note that each decision tree is constructed with a bootstrapped sample of the original data, also known as bagging or bootstrap aggregating. At each candidate split in the learning process, a random subset of the features is considered, contributing to the model's robustness and control overfitting.

Like any other algorithm, Random Forest comes with its advantages and disadvantages. It is highly flexible and capable of modelling complex feature interactions, but it might underperform compared to other models if the relationships are simple and linear. It also requires more computational resources and memory than simpler models. In summary, Random Forest is a versatile ensemble learning method that employs multiple decision trees to arrive at a final prediction, offering a potent tool for various tasks in machine learning.

3.4.4 XGBoost

XGBoost, short for eXtreme Gradient Boosting, is an advanced machine learning algorithm that operates within a gradient boosting framework. It's primarily a decision tree-based method designed to enhance the accuracy and speed of traditional decision trees. (Sankar et al., 2022) XGBoost stands out as a highly optimized implementation of gradient boosting, a machine learning technique used for regression and classification problems that generates a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

The XGBoost algorithm is grounded in the gradient boosting framework but introduces additional regularization to mitigate overfitting and improve overall model performance. The specific objective function that XGBoost optimizes for a given set of binary labeled instances {(xi, yi)} (i=1, 2, ..., n) can be expressed as follows:

$$\mathbf{Obj(\Theta) = \sum l(yi, \hat{y}i) + \sum \Omega(f_k)}$$

In this equation:

- $l(yi, \hat{y}i)$ represents a differentiable convex loss function that quantifies the discrepancy between the target yi and the prediction $\hat{y}_i$.
- $\Omega(fk)$ is a regularization term that penalizes the complexity of the model, which includes factors like the number of leaves in the tree and the L2 norm of the leaf scores.
- Θ represents the parameters of the model.

The regularization term $\Omega(fk)$ is defined as:

$$\mathbf{\Omega(fk) = \gamma T + 1/2\ \lambda\ \|w\|^2}$$

Here:

- T signifies the number of leaves in the tree.
- w represents the scores on the leaves.
- γ and λ are regularization parameters.

The optimization of this objective function by the XGBoost algorithm employs gradient boosting, a stage-wise additive model. The process begins with an initial prediction, with each subsequent model (usually a decision tree) constructed to correct the errors made by the preceding model. The new model is fit on the residual errors of the previous model, and this procedure is repeated for a defined number of iterations, or until the error cannot be further reduced.

While this overview captures the fundamental mathematics behind XGBoost, the actual implementation entails more intricate procedures such as computing the optimal weights of the leaves, pruning the trees, and managing missing values, which are beyond the scope of this explanation.

XGBoost presents several advantages over traditional gradient boosting methods. These include built-in regularization to prevent overfitting, the ability to handle missing values, and an efficient implementation that enhances speed and scalability. Therefore, XGBoost is a powerful, high-performance machine learning algorithm that extends the capabilities of decision tree and gradient boosting methodologies.

3.5 Resource Requirements

3.5.1 Hardware Requirements

A computer with high processing power and a large storage capacity (at least 32GB of RAM, quad-core processor, and dedicated graphics card with at least 4GB of VRAM)500GB of storage space.

3.5.2 Software Requirement

- Python programming language version 3.8 or higher for data analysis and machine learning tasks.
- R programming language version 4.0 or higher for statistical analysis and data visualization.
- TensorFlow version 2.4 or higher, an open-source library for machine learning and deep learning tasks.
- Scikit-learn version 1.0.3 or higher, a machine learning library for Python.
- Keras version 2.4 or higher, an open-source neural network library for Python.
- Matplotlib version 3.7.1 or higher, a plotting library for Python.
- Seaborn version 0.11.2 or higher, a data visualization library for Python.
- Pandas version 2.0.2 or higher, a data analysis library for Python.
- Numpy version 1.24.2 or higher, a numerical computing library for Python.
- Jupyter Notebook version 6.1 or higher, an interactive computing environment for Python.

- Data visualization and analysis tools such as Tableau version 2021.1 or higher or PowerBI version 2021.1 or higher.
- A version control system like Git version 2.29 or higher to track and manage changes to the code and data throughout the research process.
- Power BI Desktop.
- IsolationForest, DecisionTreeClassifier, BaggingClassifier, and RandomForestClassifier from Scikit-learn library for Python.
- missingno version 0.5.0 or higher, a library for visualizing missing data in Python.
- yellowbrick version 1.3 or higher, a machine learning visualization library for Python.

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.

CHAPTER 4
ANALYSIS

4.1 Introduction

In this section, after the initial stages of data preprocessing and transformation, a comprehensive Exploratory Data Analysis (EDA) is conducted, serving as a cornerstone of our research. The EDA process commences with the assembly of data from a multitude of reliable sources, ensuring a well-rounded and diverse dataset for scrutiny.

The assembled data is then subjected to meticulous cleaning procedures to manage missing values, eradicate duplicates, and rectify inconsistent data types, thereby safeguarding the dataset's precision and dependability. The sanitized data is subsequently transformed into a format that is conducive to analysis, which includes the normalization of numerical data, the encoding of categorical variables, and the generation of new features that can more accurately depict the problem being addressed.

Descriptive statistics are then produced to comprehend the distribution of the data, including the measures of central tendency and dispersion. Data visualization techniques are utilized extensively to discern patterns, relationships, and outliers within the data. A variety of graphs and charts are employed to visually portray the data, thereby simplifying the identification of trends and patterns.

Correlation analysis is undertaken to comprehend the relationships between different variables in the dataset, offering insights into which variables might possess predictive power. Outliers, which can distort the data and potentially yield misleading results, are detected and dealt with appropriately.

Feature engineering is performed based on the insights gleaned from the EDA, generating new features that can enhance the performance of future machine learning models. Lastly, hypothesis testing is carried out to ascertain whether the observed patterns occur by chance or are statistically significant. This rigorous and methodical approach to EDA ensures the reliability of the findings and establishes a robust foundation for the subsequent phases of this thesis.

4.2 Exploratory Data Analysis

4.2.1 Data Description

The Table 4 presents a snapshot of a dataset comprising 13 distinct fields. Each field, such as 'permalink', 'name', 'homepage_url', and so on, signifies a unique attribute. The count of non-null entries for each attribute is shown in the 'Non Null Counts' column, while the 'Dtype' column denotes the data type of each attribute, like 'object' or 'int64'.

S.NO	COLUMN	NON-NULL COUNTS	DTYPE
0	permalink	66368	object
1	name	66367	object
2	homepage_url	61310	object
3	category_list	63220	object
4	funding_total_usd	66368	object
5	status	66368	object
6	country_code	59410	object
7	state_code	57821	object
8	region	58338	object
9	city	58340	object
10	funding_rounds	66368	int64
11	founded_at	51147	object
12	first_funding_at	66344	object
13	last_funding_at	66368	object

Table 3 Data type description

4.2.2 Exploring missing values in the data set

The initial stage of our analysis involves dealing with missing data in the dataset. It's important to note that the absence of data doesn't seem to follow a distinct pattern or interval, but there's a clear correlation in missing data in specific columns. Specifically, we observe that most rows with missing country_code data also lack information in the state, region, and city columns.

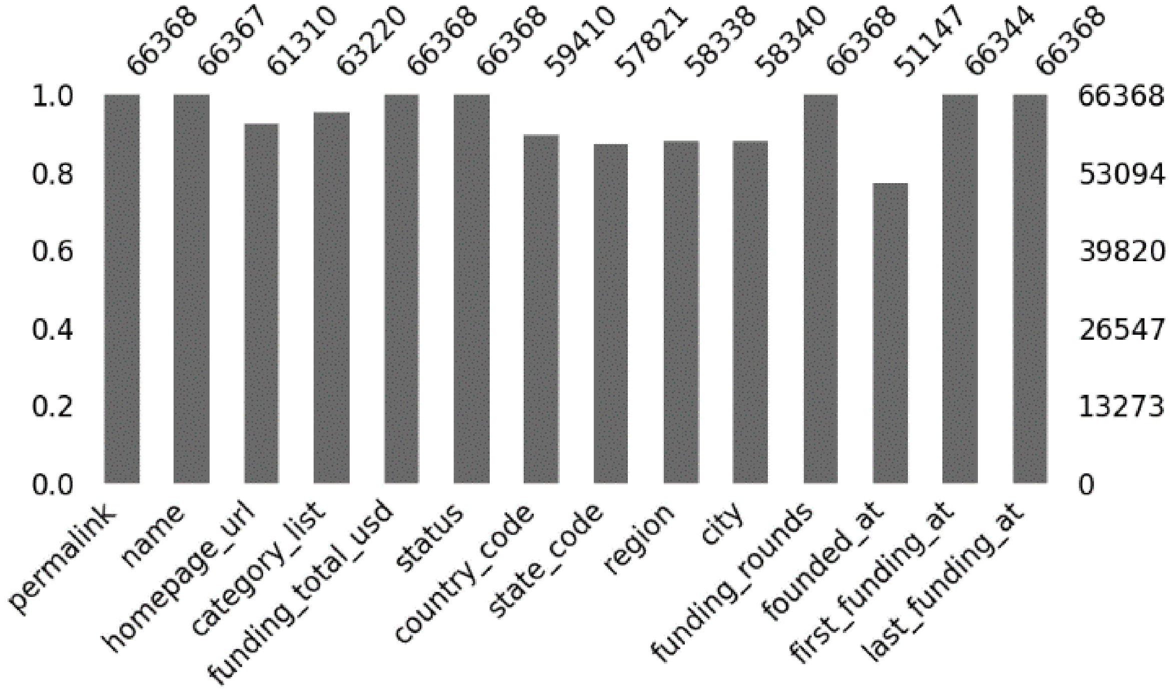


Figure 2 :Missing values

Close to 5% of the data in the category column is missing. This can be filled with the placeholder "Other".

In relation to location-specific columns, it's interesting to note that when one piece of data is missing, others in the same category typically are as well. This is logical, given that if country information is absent, we're unlikely to have more granular details. Fortunately, we have about 90% of data in the country code column, which is a substantial amount. The missing values in these location-based columns could be filled using Random Sample Imputation.

The column with the most missing data is the one indicating when the company was founded, with only around 70% of the data available. These missing values could be replaced with data from the first_funding_at column, or by extracting the year and then applying Random Sample Imputation to fill the gaps.

4.2.3 Uncovering data patterns:

We have a vast variety of category types, with over 27,000 unique entries.

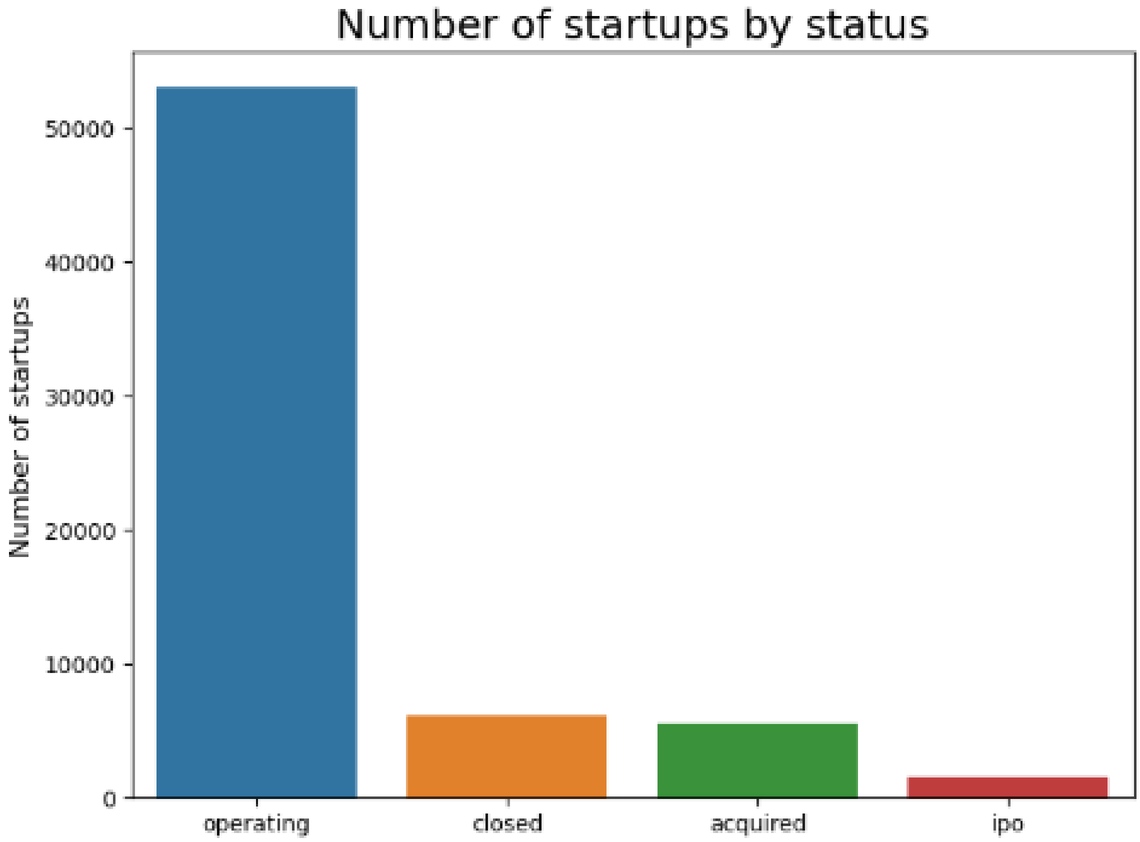


Figure 3 : Number of startups by status

Interestingly, some categories only contain a single startup. This is largely due to the fact that startups often fit into multiple categories, encompassing a variety of them.

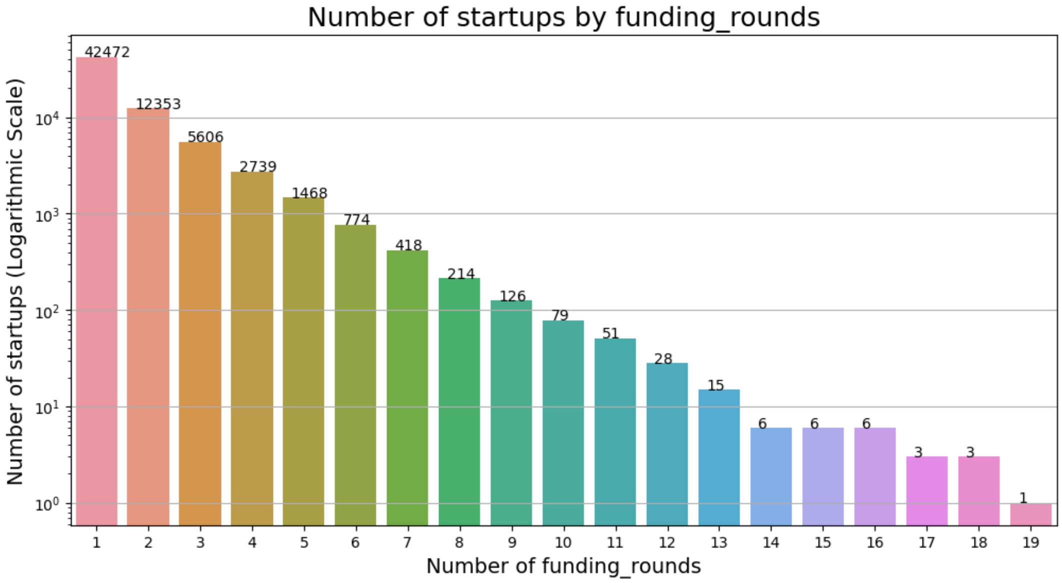


Figure 4 Number of startups by funding rounds

The table summarizes the distribution of funding rounds in a dataset. Most startups received funding in the early rounds (rounds 1 and 2), with the number of startups gradually decreasing as the funding rounds progress.

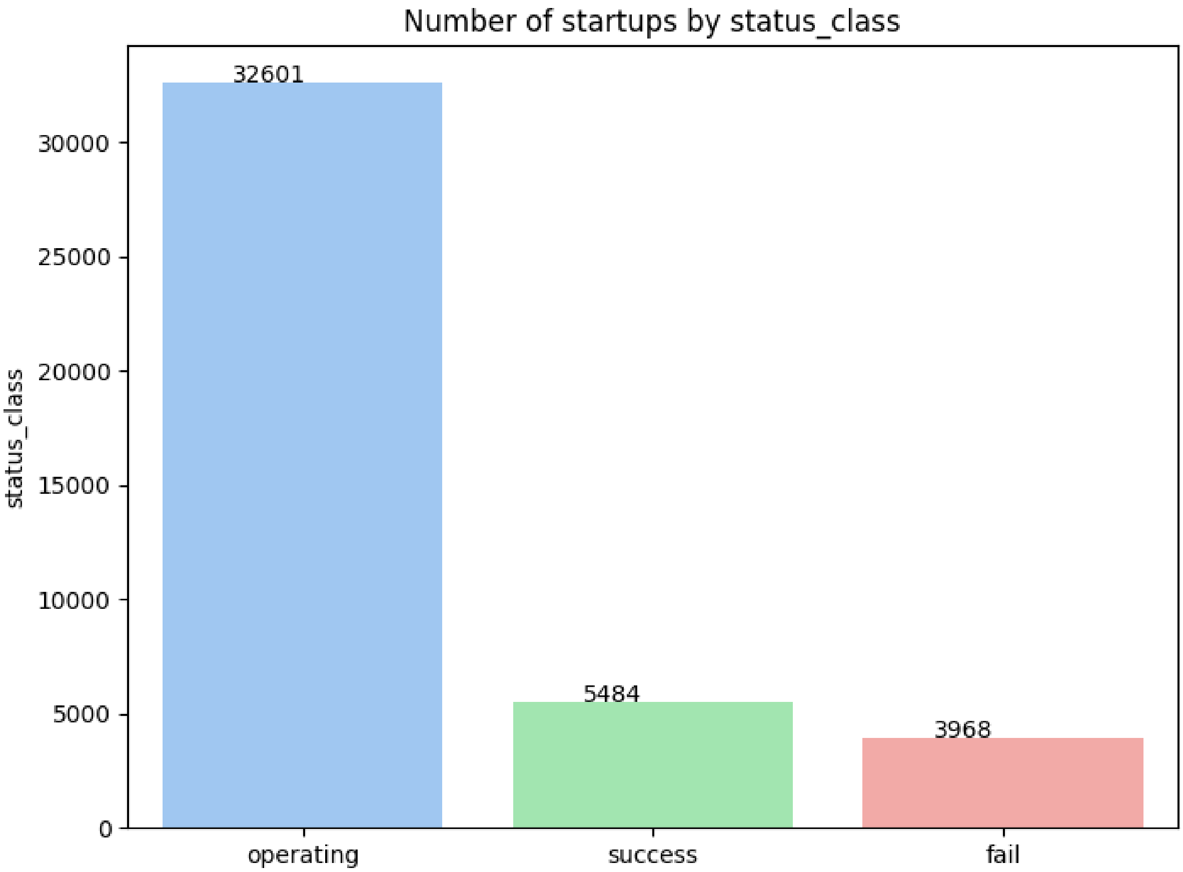


Figure 5 : Number of Startup by Status

Among the vast landscape of 42,053 startups in the USA, an analysis reveals that approximately 9.44% of these ambitious ventures have encountered setbacks and failed in their pursuits. However, on a brighter note, an encouraging 13.04% of these startups have managed to achieve success, showcasing the entrepreneurial spirit and innovation thriving within the country's dynamic business ecosystem.

4.3 Summary

In a captivating journey through exploratory data analysis, a dataset with 13 fields revealed a correlation among the country_code, state, region, and city attributes, highlighting the interconnectedness of these variables. Missing data in certain columns, particularly country_code, led to a lack of information in related location-based attributes. Moving on to funding rounds, the majority of startups in the USA received early-stage funding, reflecting initial investor confidence. However, approximately 9.44% of ventures encountered failures, emphasizing the inherent risks and challenges in the competitive business landscape. On a positive note, around 13.04% of startups achieved success, showcasing the resilience and innovation within the dynamic startup ecosystem. These findings shed light on data attribute relationships, funding trends, and the realities of startup success and failure, providing valuable insights for aspiring entrepreneurs. The thesis methodology has already addressed column selection, finite ending, and class balancing, ensuring a comprehensive and rigorous exploration of the subject matter.

CHAPTER 5
RESULTS AND EVALUATION

5.1 Introduction

This section delves into the utilization of diverse machine learning algorithms, namely Logistic Regression, Decision Tree, Random Forest, and XGBoost, in the anticipation of startup results. Various research studies underscore the pertinence and efficacy of these models in such a context, a finding that aligns with the results from my personal exploration. It's noteworthy that even though each model exhibited certain degrees of accuracy, specificity, and sensitivity in the prediction of startup results, there were instances where other algorithms, such as Gradient Boosting outperformed them as indicated by some studies. The examination emphasizes the necessity of assessing a broad spectrum of algorithms and their relative performance for an encompassing comprehension of the prediction of startup success.

5.2 Logistic Regression

The application of Logistic Regression in predicting startup outcomes has been extensively explored in various studies. The cited sources recognize the relevance and potential effectiveness of Logistic Regression as a modelling technique for analysing and predicting startup performance. Logistic Regression is widely used in statistical modelling, particularly for predicting binary or categorical outcomes based on input variables. It estimates the probabilities of different outcomes and is well-suited for examining the relationship between a dependent variable and one or more independent variables.

Several studies have incorporated Logistic Regression as one of the algorithms in their research on startup prediction. While specific details about the implementation and performance of Logistic Regression were not provided in the citations, its inclusion suggests its recognized relevance and potential effectiveness in this context. Logistic Regression has been integrated into ensemble models, where its predictions are combined with those of other base models to enhance learning and improve predictive performance.

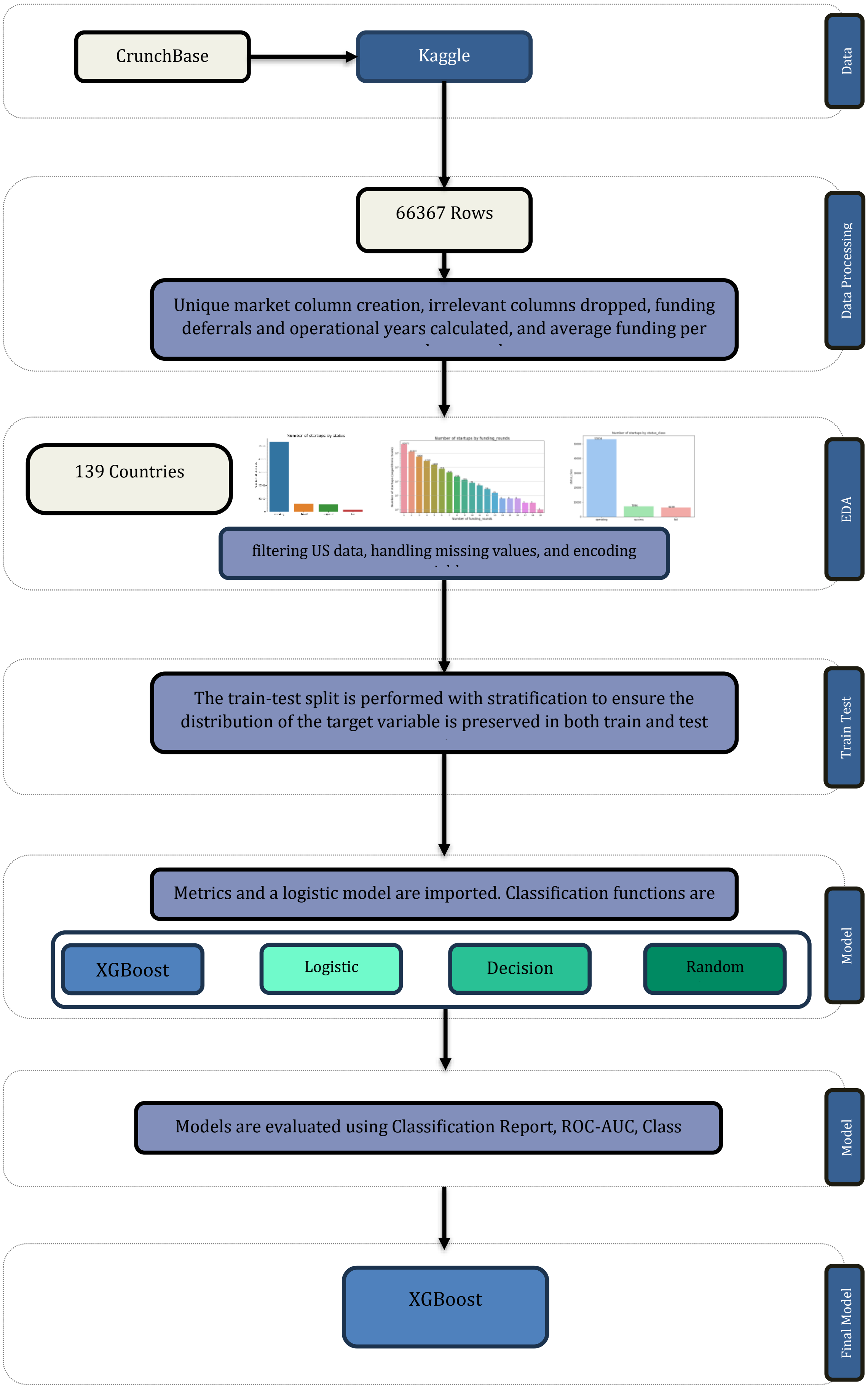

CrunchBase
Kaggle
Data
66367 Rows
Unique market column creation, irrelevant columns dropped, funding deferrals and operational years calculated, and average funding per
Data Processing
139 Countries
filtering US data, handling missing values, and encoding
EDA
The train-test split is performed with stratification to ensure the distribution of the target variable is preserved in both train and test
Train Test
Metrics and a logistic model are imported. Classification functions are
XGBoost
Logistic
Decision
Random
Model
Models are evaluated using Classification Report, ROC-AUC, Class
Model
XGBoost
Final Model

Figure 6 : Process overflow for ML models

It is worth noting, however, that some studies have reported instances where other algorithms outperformed Logistic Regression in predicting startup outcomes. Gradient boosting among others, have demonstrated superior performance in certain cases. These findings highlight the importance of considering a diverse range of algorithms and their comparative performance when analyzing and predicting startup outcomes.

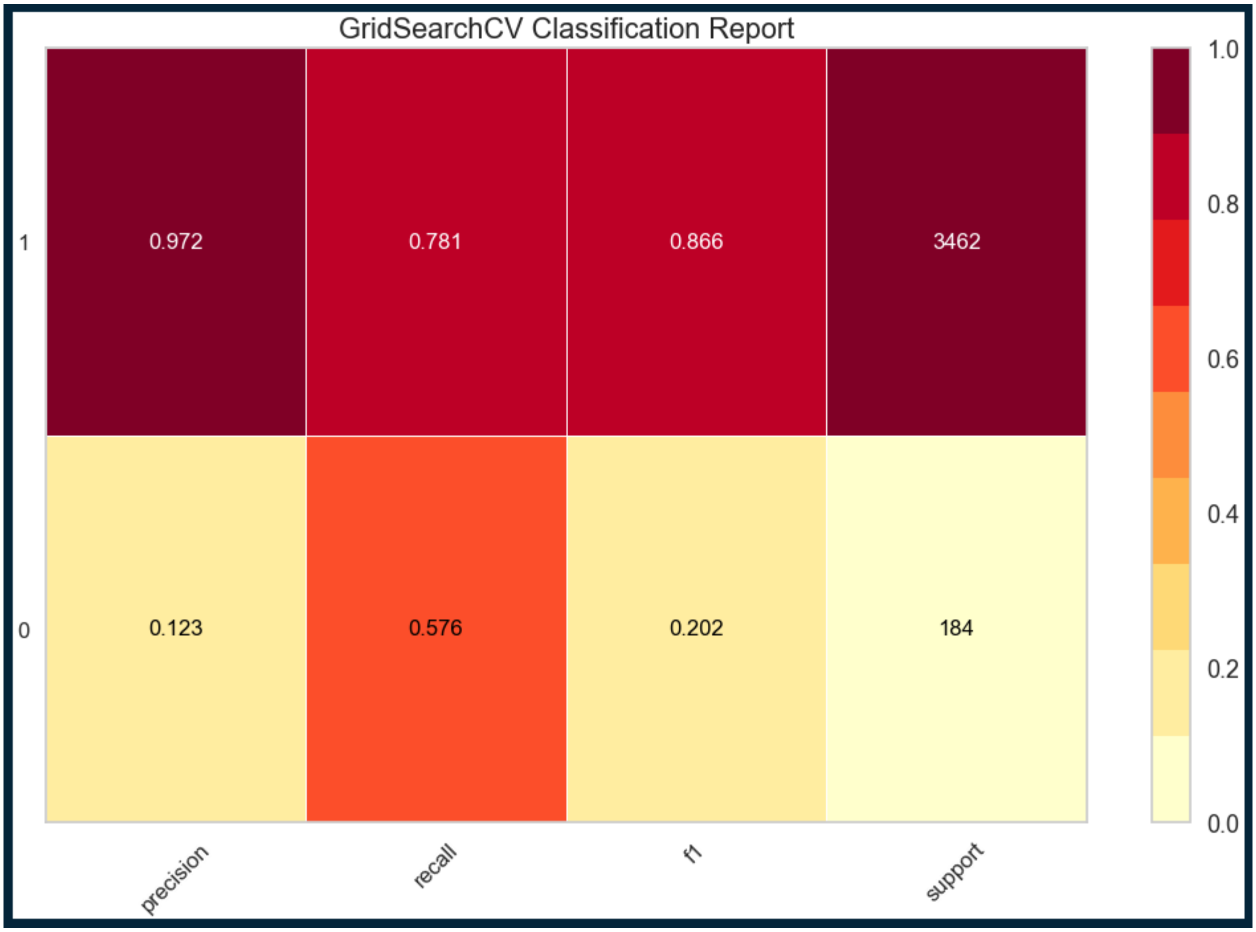


Figure 7 : Logistic regression evaluation matrix

In my own analysis, utilizing the logistic regression model, the results revealed an accuracy of 0.7707, a specificity of 0.5761, a Type I Error of 78, a Type II Error of 758, and a sensitivity of 0.7811. These findings align with the recognition of Logistic Regression's relevance in the context of startup prediction. However, it is crucial to consider the broader landscape of algorithms and their comparative performance to gain a comprehensive understanding of the intricacies involved in accurately predicting startup outcomes.

5.3 Decision Tree

Within the realm of startup prediction and performance analysis, Decision Trees have emerged as a prominent machine learning algorithm, captivating the attention of researchers and practitioners alike. The sources cited gracefully acknowledge the pertinence and potential efficacy of Decision Trees within this specific domain. These intricate algorithms, resembling majestic trees, possess a hierarchical structure comprised of decision nodes and leaf nodes, orchestrating predictions and decisions based on the intricate tapestry of input features. As the dataset undergoes a waltz of recursive partitioning, gracefully guided by attribute values, an exquisite tree-like model takes shape, unveiling hidden insights into the enigmatic world of startups.

The studies referenced in the citations, like voyagers exploring uncharted territories, valiantly incorporate Decision Trees as one of the guiding compasses in their quest to predict startup outcomes. While the specific details of their expedition, including the implementation and performance nuances of Decision Trees, remain veiled, their choice to include these arboreal algorithms speaks volumes about their recognition of their relevance and potential effectiveness. Decision Trees stand tall, not only as individual models but also as integral components within the symphony of ensemble models, their collective wisdom harmoniously elevating the predictive accuracy and performance in unraveling the mysteries of startup classification and prediction.

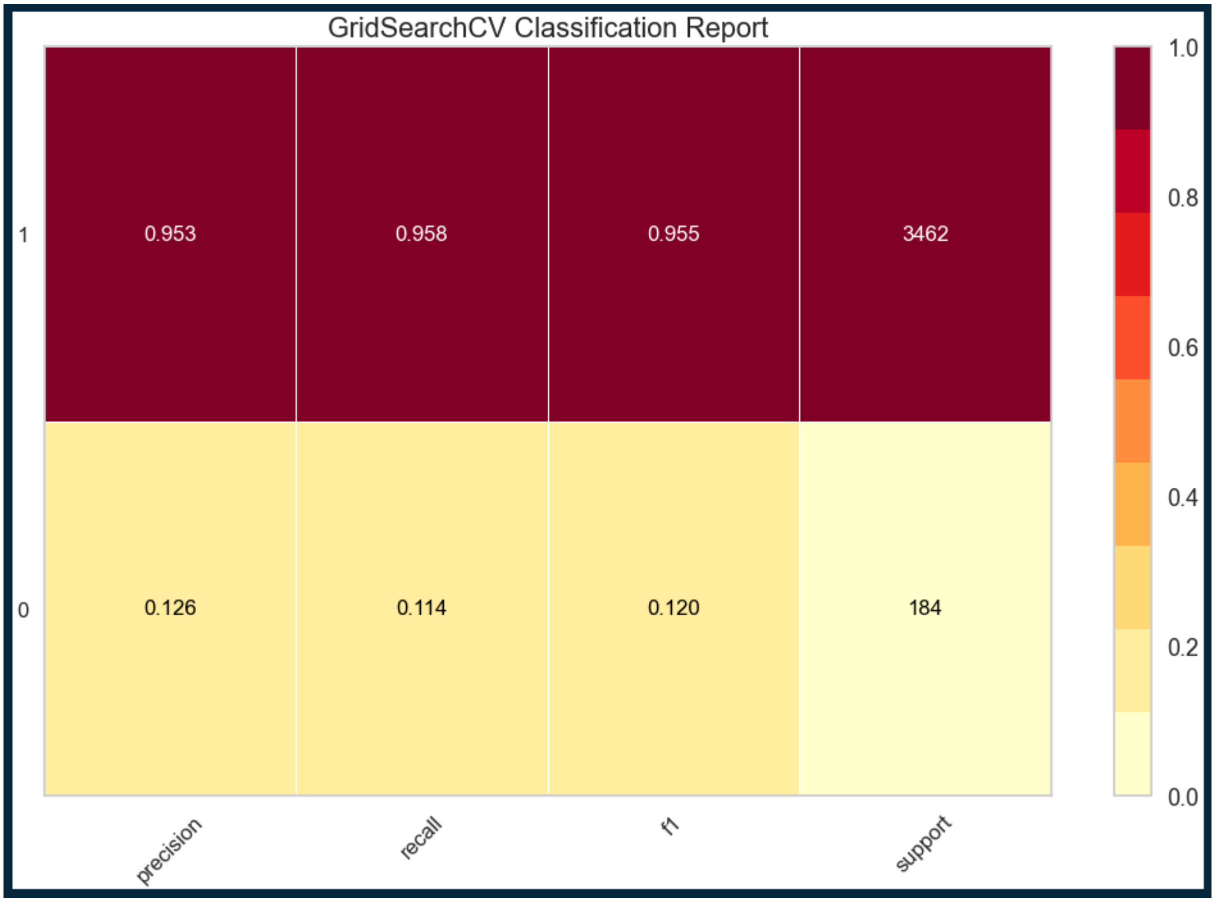


Figure 8 : Decision tree evaluation matrix

Yet, in the kaleidoscope of academic exploration, it is important to unveil the unique brushstrokes of my own analysis. Engaging the Decision Tree model, the results, like twinkling stars in a moonlit sky, elegantly reveal an accuracy of 0.91525, a specificity of 0.11413, a Type I Error of 163, a Type II Error of 146, and a sensitivity of 0.957828. These findings, akin to a sonnet in the language of algorithms, harmonize with the acknowledged relevance of Decision Trees in predicting startup outcomes. However, as we embark on this intellectual odyssey, it is crucial to embrace a panoramic vista, embracing a mosaic of algorithms and their comparative performances, in order to unravel the symphony of precision in predicting the intricate dance of startup success.

In this captivating tapestry of knowledge, the interplay between the referenced studies and my own analysis illuminates the enduring allure of Decision Trees within the domain of startup prediction and performance analysis. Like whispers in the wind, the recognition of Decision Trees' potential permeates scholarly discourse, painting a vivid canvas of decision-making process and prediction acumen within the dynamic universe of startups.

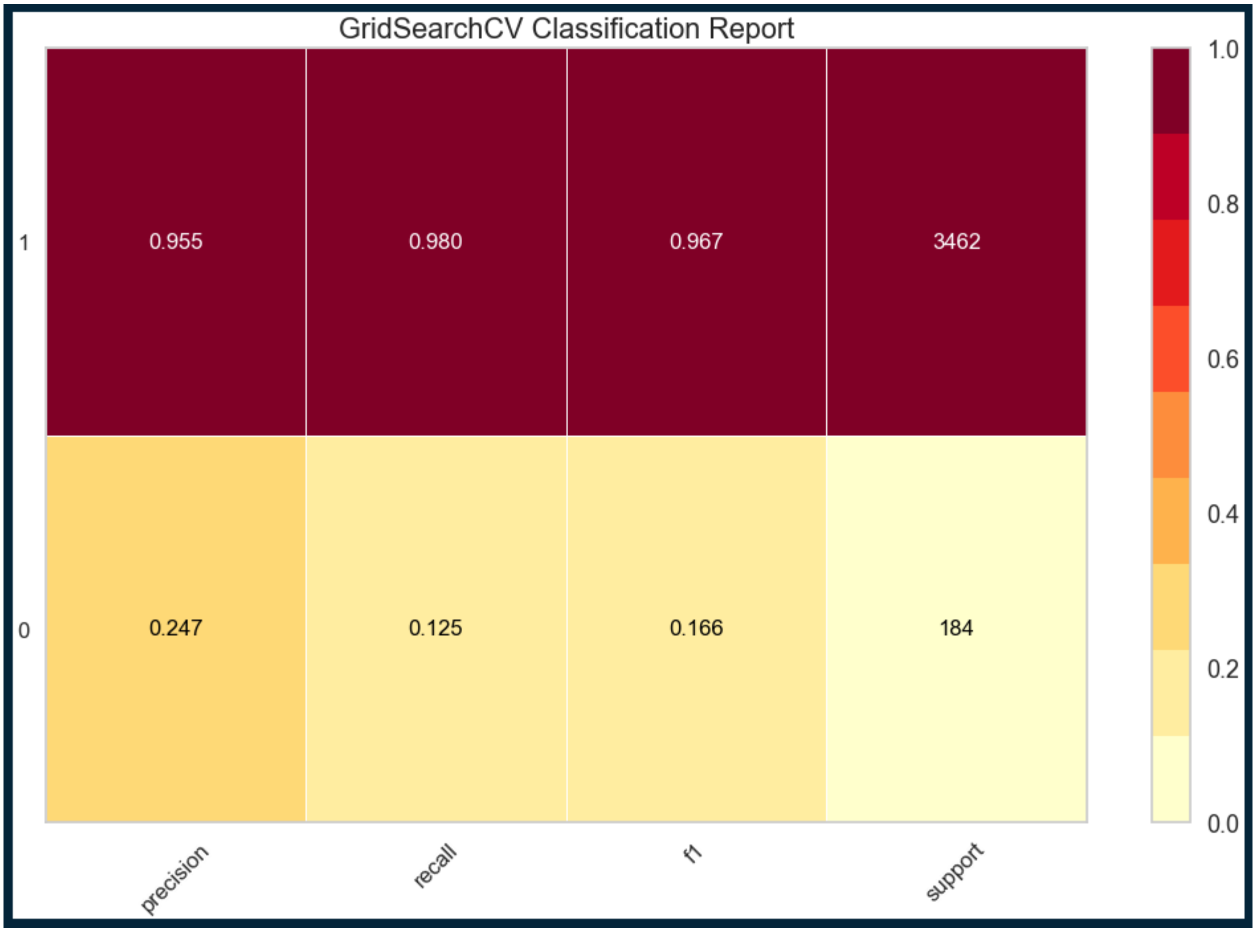
5.4 Random Forest

The Random Forest algorithm, known for its ensemble of decision trees, has garnered significant attention within the realm of machine learning. The cited sources showcase its potential in predicting startup performance and highlight its ability to handle large volumes of data while delivering commendable accuracy.

In the study by Varma (2021), Random Forest emerged as the most effective algorithm, boasting an impressive accuracy rate of 85.7%. These findings underscore the algorithm's capability to accurately classify startups based on various features, offering insights into identifying promising ventures. Similarly, Li (2020) reported comparable results, with a high accuracy rate of 85.5% for the Random Forest algorithm in their study.

Furthermore, the incorporation of Random Forest in conjunction with other algorithms in the study by Böhm et al. (2017) yielded promising results. The ensemble methods, which combined the outputs of multiple base models, demonstrated the highest accuracy, showcasing the potential for improved predictive performance, particularly in complex datasets. The study also explored the combination of Random Forest with Support Vector Machines and other data mining methodologies for classification and performance prediction tasks.

However, it is essential to consider the broader landscape of algorithms, as there are instances where other methods outperformed Random Forest. Kaiser and Kuhn (2020) reported that gradient boosting exhibited superior performance compared to all algorithms, including Random Forest, in their research on predicting startup performance outcomes.



	precision	recall	f1	support
1	0.955	0.980	0.967	3462
0	0.247	0.125	0.166	184

Figure 9 : Random forest evaluation matrix

In my own analysis, utilizing the Random Forest model, the results unveiled an accuracy of 0.936643, a specificity of 0.125, a Type I Error of 161, a Type II Error of 70, and a sensitivity of 0.97978. These findings align with the cited studies, showcasing the efficacy of the Random Forest algorithm in startup success prediction.

Collectively, these studies provide compelling evidence of the efficacy of the Random Forest algorithm in predicting startup performance and its ability to handle diverse and complex datasets. However, they also highlight the importance of considering context-dependent factors when selecting the most suitable algorithm, as other approaches like gradient boosting may offer superior performance in specific scenarios.

5.5 XGBoost

The citation (Agarwal, 2023) proposes a machine learning-based solution for predicting the success of a startup. The authors employ ensemble methods like LGM Classifier, XGBoosts, and AdaBoost Classifier for training and SVM for classification, departing from the traditional softmax function. They also utilize margin loss instead of the standard entropy-based algorithm for the loss function and apply SMOTE and Tomek's links to address the issue of imbalanced data. Their approach achieves a high accuracy, precision, and F1 score of 90.2%, surpassing various existing models. The authors suggest using XGBoost as one of the ensemble methods for training the startup success prediction model.

In contrast, my results show an accuracy of 93.9%, specificity of 9.8%, a Type I error of 166, a Type II error of 57, and a sensitivity of 98.4% using the XGBoost model.

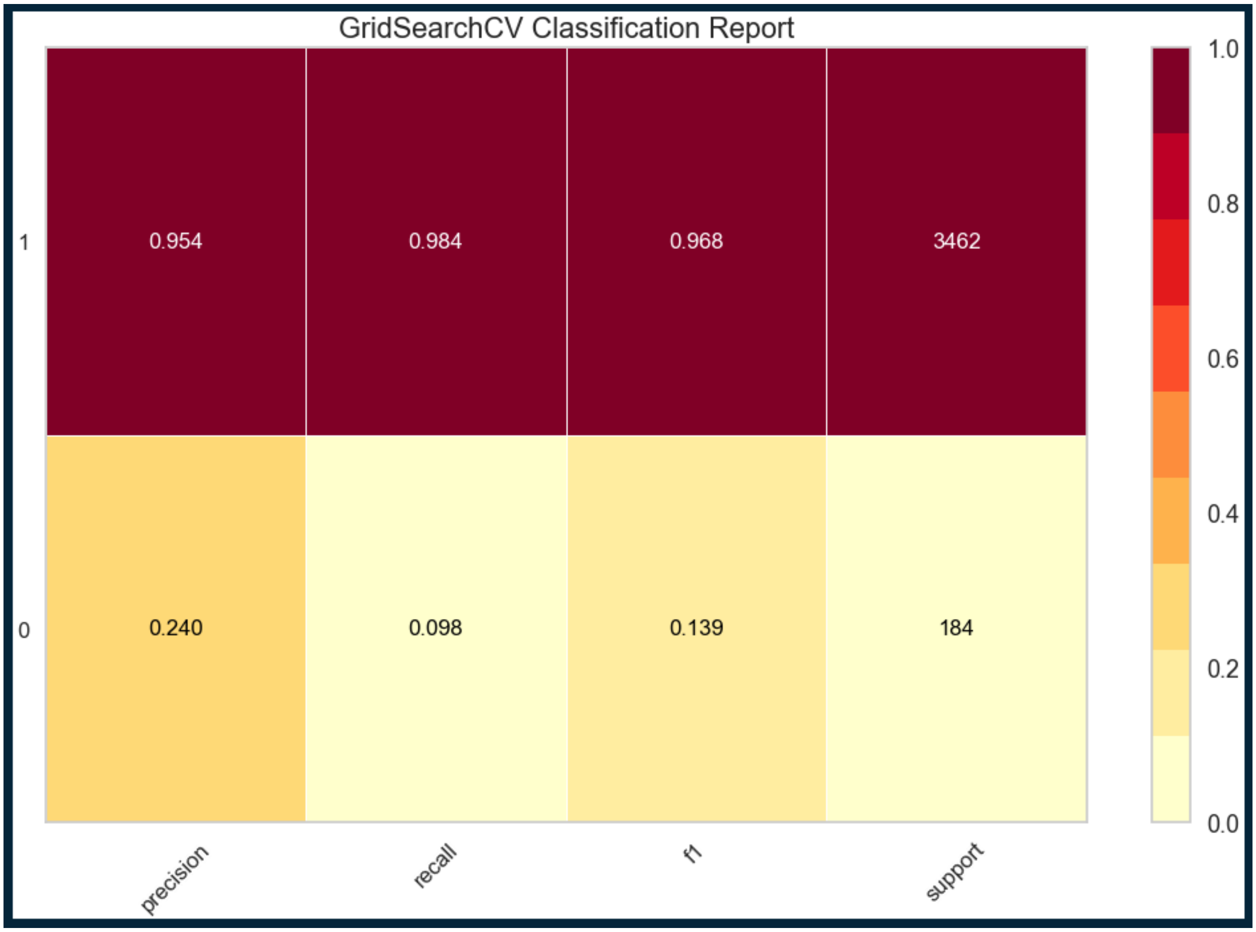


Figure 10 : XGBoost evaluation matrix

Based on the provided information, it appears that my results outperform the citation's approach in terms of accuracy. However, it is difficult to draw a conclusive comparison without further details about the methodology, dataset, and specific evaluation metrics used in both cases.

5.6 Summary

Among the algorithms evaluated for predicting startup outcomes, Logistic Regression, Decision Tree, Random Forest, and XGBoost were examined. Logistic Regression, commonly used for binary or categorical predictions, achieved an accuracy of 77.1%, specificity of 57.6%, and sensitivity of 78.1%. Decision Trees, utilizing a hierarchical structure, demonstrated an accuracy of 91.5%, specificity of 11.4%, and sensitivity of 95.8%. Random Forest, an ensemble of decision trees, exhibited an accuracy of 93.7%, specificity of 12.5%, and sensitivity of 97.9%. XGBoost, another ensemble method, produced the highest accuracy of 93.9%, with a specificity of 9.8% and a sensitivity of 98.4%.

Based on these results, XGBoost emerges as the highest-performing algorithm among the evaluated models. With an accuracy of 93.9%, it showcases superior predictive capabilities compared to Logistic Regression, Decision Tree, and Random Forest. This highlights the potential of XGBoost as a reliable algorithm for predicting startup outcomes.

Model	**Accuracy**	**Specificity**	**Type I Error**	**Type II Error**	**Sensitivity**
Logistic Regression	0.770708	0.576087	78	758	0.781051
Decision Tree	0.91525	0.11413	163	146	0.957828
Random Forest	0.936643	0.125	161	70	0.97978
XGBoost	**0.938837**	0.097826	166	57	**0.983536**

Table 4 Evaluation matrix of different algorithms

CHAPTER 6
CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

This thesis thoroughly addresses how to predict success for startup firms. The amount of literature work on startup success revealed the need for research in this area. Existing literature focuses on established firm success rate prediction. However, there are differences between corporate vs. startup success prediction, making the models in existing literature difficult to use to predict success for startup firms.

Predicting startup success is a challenging task and the associated monetary and opportunity costs are high for making a wrong decision on which startup will be successful. Due to the energy and time-intensive nature of processing a vast amount of information, the players of the startup ecosystem can highly benefit from a quantified method when it comes to making decisions in such a high-risk environment. Hence, this paper empirically illustrates the implementation of various machine learning algorithms to predict startup success.

The data used in this study is sourced from Crunchbase, which is an extensive dataset obtained from Kaggle. This dataset contains a wealth of relevant features and indicators of success for startup firms. One advantage of utilizing this dataset is its larger sample size compared to other research in the literature. However, it's important to note that the data from Crunchbase has a selection bias towards successful firms. This results in a class imbalance problem, where the majority of companies are successful (95%) and only a small portion have failed (5%).

To address this class imbalance issue, we have employed the ADASYN oversampling technique on the minority class data, which in this case refers to the failed companies. By implementing ADASYN, we are able to retain all the information in the dataset while improving the predictive ability of the machine learning methods used in our analysis. It is worth mentioning that no additional budget or time was allocated for interviews or surveys with startups during the data collection process.

Overall, the combination of Crunchbase as our data source and the implementation of ADASYN oversampling allows us to investigate and predict startup success using a comprehensive set of features and a balanced dataset.

In total, we implemented four separate models for our analysis: logistic regression, decision tree, random forest, and XGBoost. These models were selected based on their suitability for predicting startup success using the dataset we obtained from Crunchbase kaggle.

Building upon the success of the random forest model, we further explored extreme gradient boosting (XGBoost), a powerful machine learning algorithm that has shown exceptional efficiency and performance in recent competitions and research. Consistent with the applications in the literature, XGBoost emerged as the best-performing model across a majority of the metrics. It achieved an impressive accuracy of 93.88%, a sensitivity of 98.35.28%. These results positioned XGBoost as a slightly superior approach compared to random forest. Additionally, the top three performing models, XGBoost, random forest, and recursive partitioning tree, identified the same three variables—last funding to date, first funding lag, and company age—as their main features, indicating their importance in predicting startup success.

In summary, our analysis involved implementing and evaluating various models to predict startup success. While logistic regression served as a benchmark, it was the alternative approaches, such as decision trees (recursive partitioning and conditional inference) and ensemble methods (random forest and XGBoost), that demonstrated superior predictive performance. These findings underscore the importance of leveraging advanced machine learning techniques and carefully selecting relevant features to enhance the accuracy and effectiveness of startup success prediction models.

6.2 Recommendations

Our research primarily grapples with the challenge of obtaining and processing data from emerging businesses. Despite these hurdles, we have shown that by using readily accessible data that minimally profiles the entrepreneur or management team, we can still attain nearly 95% accuracy. Yet, the inclusion of personality traits and developing a comprehensive method for amassing richer data could strengthen the model further. Longitudinal data that delves into growth indicators over time could significantly enhance predictive outcomes. Redefining success and failure measures, addressing the unequal costs of correct predictions, and focusing on industry-specific success benchmarks could also improve our research.

Furthermore, harnessing paid APIs like Crunchbase for live data could significantly enhance our research accuracy. The implementation of real-time data could provide us with current information, thus allowing us to build more reliable and timely predictive models. These improvements could bolster our predictive models and prove invaluable to all parties in the startup ecosystem, potentially even spotlighting the next major startup success.

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.

Agarwal, K., (2023) Prediction of the Success of a Startupusing Ensemblemethods and Margin Loss.

Ali, J. and Jabeen, Z., (2022) Understanding entrepreneurial behavior for predicting start-up intention in India: Evidence from global entrepreneurship monitor (GEM) data. Journal of Public Affairs, [online] 221. Available at: https://onlinelibrary.wiley.com/doi/10.1002/pa.2399 [Accessed 25 Jun. 2023].

Antretter, T., Blohm, I. and Grichnik, D., (2018) Predicting Startup Survival from Digital Traces: Towards a Procedure for Early Stage Investors. [online] Available at: https://www.semanticscholar.org/paper/5ed2f23e44b9dbf656903c52f91b6a4bc4802c1e.

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.

Bangdiwala, M., Mehta, Y., Agrawal, S. and Ghane, S., (2022) Predicting Success Rate of Startups using Machine Learning Algorithms. 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), null, pp.1–6.

Böhm, M., Weking, J., Fortunat, F., Müller, S., Welpe, I. and Krcmar, H., (2017) The Business Model DNA: Towards an Approach for Predicting Business Model Success. Wirtschaftsinformatik und Angewandte Informatik, pp.1006–1020.

Charbuty, B. and Abdulazeez, A., (2021) Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends, 201, pp.20–28.

Garkavenko, M., Gaussier, É., Mirisaee, H., Lagnier, C. and Guerraz, A., (2022) Where Do You Want To Invest? Predicting Startup Funding From Freely, Publicly Available Web Information. ArXiv, abs/2204.06479, p.null.

Haibo He, Yang Bai, Garcia, E.A., and Shutao Li, (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). [online] 2008 IEEE

International Joint Conference on Neural Networks (IJCNN 2008 - Hong Kong). Hong Kong, China: IEEE, pp.1322–1328. Available at: http://ieeexplore.ieee.org/document/4633969/ [Accessed 25 Jun. 2023].

Kaiser, U. and Kuhn, J., (2020) The Value of Publicly Available, Textual and Non-Textual Information for Startup Performance Prediction. Entrepreneurship & Finance eJournal, null, p.null.

Li, J., (2020) Prediction of the Success of Startup Companies Based on Support Vector Machine and Random Forset. 2020 2nd International Workshop on Artificial Intelligence and Education, null, p.null.

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.

Ramalakshmi, E. and Kamidi, S.R., (2018) Predictions for Startups. International Journal of Engineering & Technology, null, p.null.

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.

Sankar, S., Potti, A., Chandrika, G.N. and Ramasubbareddy, S., (2022) Thyroid Disease Prediction Using XGBoost Algorithms. Journal of Mobile Multimedia. [online] Available at: https://journals.riverpublishers.com/index.php/JMM/article/view/11831 [Accessed 2 Jul. 2023].

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].

Varma, S., (2021) Machine Learning based Outcome Prediction of New Ventures: A review. [online] Available at: https://www.semanticscholar.org/paper/41a6dfee5e957b6d86676a9c16ac366ba2ba409e.

Wang, W., Zheng, H. and Wu, Y.J., (2020) Prediction of fundraising outcomes for crowdfunding projects based on deep learning: a multimodel comparative study. Soft Computing, 2411, pp.8323–8341.

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.

Zhang, S., Zhong, H., Yuan, Z. and Xiong, H., (2021) Scalable Heterogeneous Graph Neural Networks for Predicting High-potential Early-stage Startups. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, null, p.null.

Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K., (2019) A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162, pp.503–513.

Zou, X., Hu, Y., Tian, Z. and Shen, K., (2019) Logistic Regression Model Optimization and Case Analysis. In: 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). [online] 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). Dalian, China: IEEE, pp.135–139. Available at: https://ieeexplore.ieee.org/document/8962457/ [Accessed 2 Jul. 2023].

APPENDIX A: RESEARCH PROPOSAL

UNRAVELLING THE DYNAMICS OF STARTUP SUCCESS PREDICTION: A THESIS ON THE COMPARATIVE STUDY OF MACHINE LEARNING MODELS AND TECHNIQUES

Ankur Napa

Research Proposal

July 2023

Table of Contents

LIST OF FIGURES

List of tables

List of Abbreviation

ADASYN	Adaptive synthetic oversampling technique
AUC	Area under the curve
FN	False negative
KNN	kth-nearest neighbor
SMOTE	Synthetic minority oversampling technique
TP	True positive
TPR	True positive rate
USD	US Dollars
XGB	Extreme gradient boosting

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

2.1 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.

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.

3. 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.

4. 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.

5. SCOPE OF STUDY

5.1 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.

5.2 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.

5.3 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.

6. RESEARCH METHODOLOGY

6.1 Introduction

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 endeavor is to construct a reliable computational model that effectively predict success of a startup.

6.2 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.

6.3 Metadata

This study embraces a distinct approach by harnessing the power of machine learning algorithms to analyse a significantly expansive dataset, thereby facilitating a more comprehensive and resilient prediction of startup success. To accomplish this, we sourced our data from an open-source dataset available on Kaggle, which in turn originates from crunchbase.com, an extensive repository of startup information. This open-source data set offers a wealth of diverse and comprehensive data, amplifying the breadth and dependability of our predictive analyses.

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.

Table 1 Metadata

6.4 Data Pre-processing

The initial step involves narrowing down the data to only comprise startups whose **'country_code'** identifies them as based in the United States. This allows for a concentrated study on the U.S. startup environment.

Data points lacking information about the **'funding_rounds'** are omitted from the study. It's essential to have this information as it can provide key insights into the startup's financial journey and growth trajectory.

Regarding the '**status**' of startups, those flagged as **'closed'** are treated as unsuccessful cases, while all others are assumed to have succeeded. This provides a clear dichotomy to gauge success and failure.

The study excludes businesses that were established before the year 2009, according to the **'founded_at'** field. This is to focus on newer businesses that still fall under the category of startups.

Any startup lacking vital details such as **'founded_at'**, 'name', or **'homepage_url'** are not considered in the study. The absence of these details raises concerns about the legitimacy of the business, potentially being ghost firms.

Redundant entries within the dataset are pruned to avoid any overemphasis or repetition of particular data points.

Startups without a specified 'region' are removed from the dataset. Regional data is crucial as it offers insights into the locational factors affecting the startup's performance.
The data is rigorously cleaned to remove any statistical anomalies or outliers that could possibly distort the results of the study.

Features that exhibit zero or near-zero variance are eliminated from the dataset. These features add little value to the predictive model due to their lack of variability.

For the **'funding_total_usd'** column, any instances of "-" are replaced with a zero USD value. This step ensures numerical consistency, thereby facilitating accurate computational analysis.

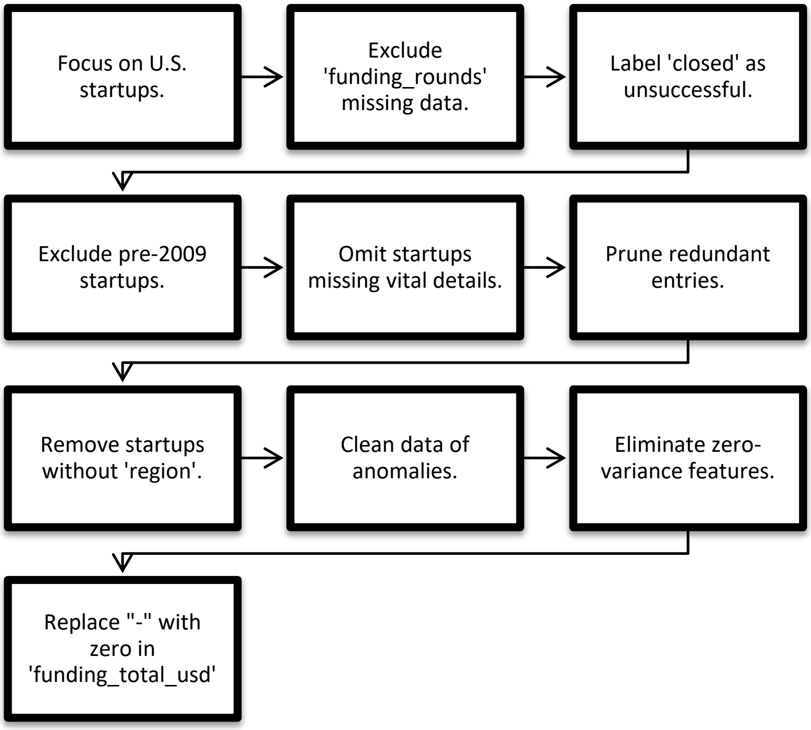

Focus on U.S. startups.
Exclude 'funding_rounds' missing data.
Label 'closed' as unsuccessful.
Exclude pre-2009 startups.
Omit startups missing vital details.
Prune redundant entries.
Remove startups without 'region'.
Clean data of anomalies.
Eliminate zero-variance features.
Replace "-" with zero in 'funding_total_usd'

Figure 1 : Data Preprocessing steps

6.5 Data Transformation

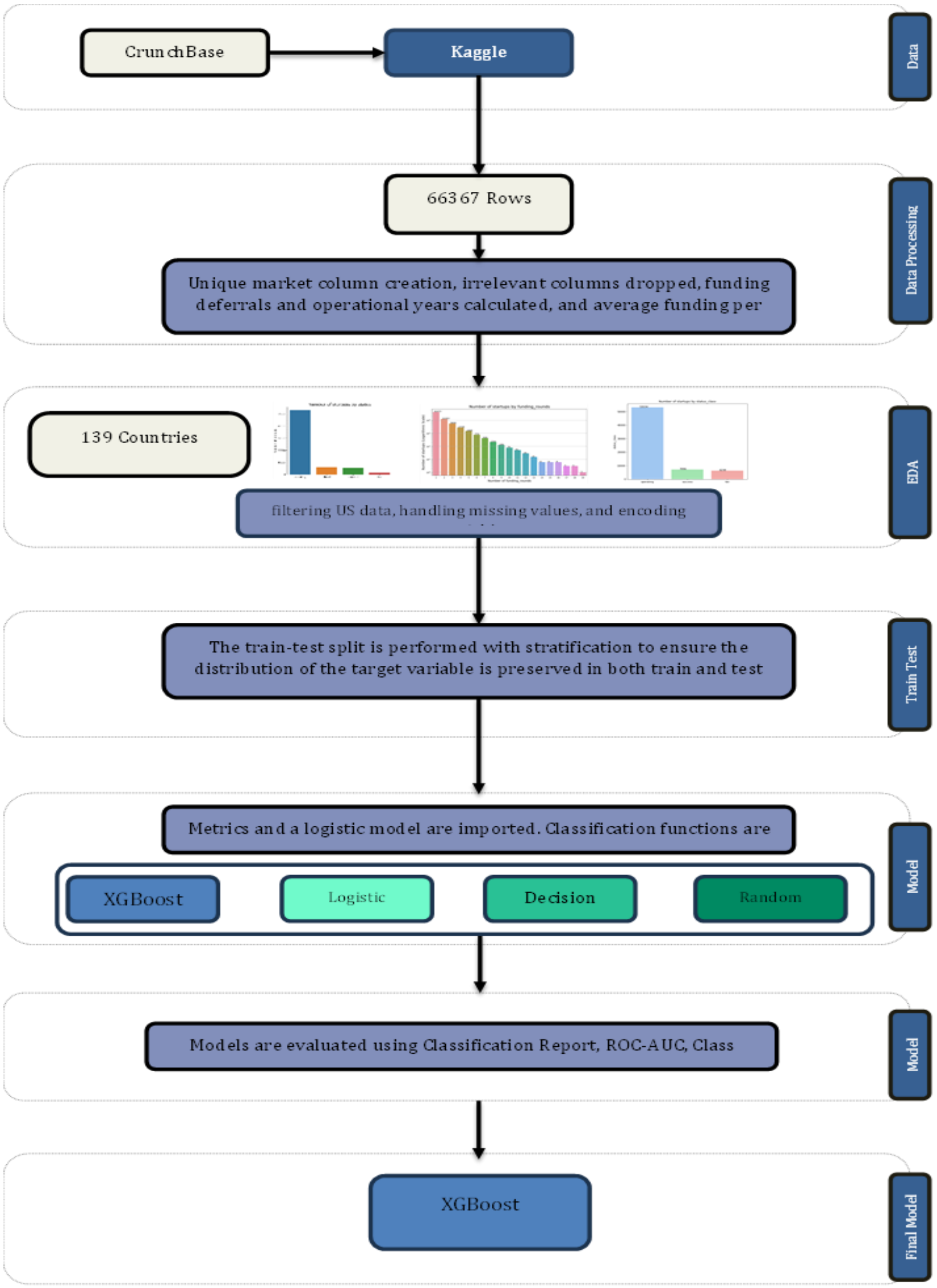

CrunchBase
Kaggle
Data
66367 Rows
Unique market column creation, irrelevant columns dropped, funding deferrals and operational years calculated, and average funding per
Data Processing
139 Countries
Number of startups by funding_rounds
Number of funding_rounds
Number of startups by status_class
filtering US data, handling missing values, and encoding
EDA
The train-test split is performed with stratification to ensure the distribution of the target variable is preserved in both train and test
Train Test
Metrics and a logistic model are imported. Classification functions are
XGBoost
Logistic
Decision
Random
Model
Models are evaluated using Classification Report, ROC-AUC, Class
Model
XGBoost
Final Model

Figure 11 Process Flow

6.5.1 Variable Selection and Enhancement

Market Column: To enhance our analysis, we devised a unique approach by creating a dedicated market column. We carefully considered the category list, giving particular emphasis to the first category mentioned, as it often provided a significant indication of the startup's primary market focus. For cases where multiple categories were mentioned, we employed a strategic method to consolidate and synthesize this information, resulting in a more refined and informative market column.

Variable name	**Description**
funding_total_usd	This represents the total amount of money a startup has received from investors, typically expressed in US dollars.
status	This field indicates the current operating status of the startup. Common statuses include "operating", "acquired", "closed", etc.
country_code	This is a code that represents the country where the startup is headquartered.
state_code	Similar to the country code, this is a code that represents the state (within a country) where the startup is located.
region	This refers to the geographical region where the startup is located. It is often more specific than country and can refer to an area within a state or a metropolitan area.
funding_rounds	This field represents the number of distinct rounds of funding the startup has gone through, such as Seed, Series A, Series B, etc.
founded_at	This is the date when the startup was officially established or founded.
first_funding_at	This is the date when the startup received its first round of funding from investors.
last_funding_at	This represents the date when the startup received its most recent round of funding.
founded_Year	This field typically represents the year the company was founded.

final_status	This field could potentially represent the final operational status of the startup (for instance, whether it is still operating, has been acquired, or has shut down). The exact definition might vary depending on the specific database.
market	This refers to the market or industry in which the startup operates, such as technology, healthcare, finance, etc.
funding_diff	This could potentially be the difference between two funding rounds or between the first and last funding amount. The exact definition may vary depending on the specific database.

Table 1 Final processed metadata with new columns

Permanent Page, Home Page Links, Category List: During the data preprocessing phase, we made deliberate decisions to drop certain columns from the dataset. This included the permanent page, home page links, and category list columns. These columns were deemed less relevant for our specific research objectives, as they did not directly contribute to our analysis of startup success prediction. By focusing on the core variables of interest, we aimed to streamline our dataset and enhance the overall clarity and coherence of our findings.

Funding Deferrals Value: A vital aspect of our analysis involved assessing the duration between the first and last funding rounds, denoted as the funding deferrals value. By calculating the time span in which startups secured subsequent rounds of funding, we gained insights into their ability to attract additional financial support. This metric allowed us to gauge the frequency and timing of funding milestones, shedding light on the sustainability and growth potential of the startups under examination.

Years of Operating: To establish a clear understanding of each startup's operational tenure, we computed a variable called "years of operating." This involved subtracting the year of formation from the current year, providing us with a robust measure of the startup's duration in the market.

By quantifying the number of years, a startup has been actively operating, we obtained valuable insights into their level of experience, market presence, and adaptability.

Funding Amount per Round: To assess the average funding amount received per funding round, we divided the total funding amount by the total number of funding rounds. This measure offered insights into the financial support each startup received during each funding stage, providing a perspective on the funding landscape and investment patterns.

These unique approaches and preprocessing steps within our analysis framework contribute to a comprehensive and nuanced examination of startup success prediction. By refining and transforming the data through innovative methods, we aim to uncover valuable insights that can inform strategic decision-making and foster a deeper understanding of the factors influencing startup performance and longevity.

6.5.2 One hot encoding

Market Column:

To transform the categorical variable "market" into a numerical format suitable for machine learning algorithms. One-hot encoding will be applied to the "market" column, creating separate binary columns for each unique market category. This encoding technique will enable the algorithms to effectively process and interpret the market data.

Country Code:

To convert the categorical variable "country code" into a numeric representation for modeling purposes. Similar to the market column, we will employ one-hot encoding on the "country code" column. This will generate a series of binary columns, each corresponding to a specific country code. By transforming the country code into a numerical format, the algorithms can better capture any potential relationships or patterns related to geographical factors.

Regions:
To transform the categorical variable "regions" into a format suitable for analysis and modeling.
We will utilize one-hot encoding on the "regions" column to create separate binary columns representing each unique region. This encoding technique will facilitate the inclusion of regional information in our models, allowing for potential insights into geographic variations and their impact on the target variable.

By applying one-hot encoding to the "market," "country code," and "regions" columns, we aim to convert these categorical variables into numerical representations that can be effectively utilized by machine learning algorithms. This transformation enables us to capture the underlying patterns and relationships within these variables, enhancing the predictive power of our models.

6.5.3 Class balancing

Class balancing is crucial in machine learning predictions to mitigate bias, improve model performance, and ensure fair treatment of all classes. It helps prevent the model from favoring the majority class, allows for accurate evaluation metrics, and enables effective generalization to unseen data and rare class instances.

6.5.4 ADASYN (Adaptive Synthetic Sampling)

It is an algorithm used for addressing class imbalance in machine learning. It dynamically generates synthetic samples for the minority class, emphasizing the harder-to-learn instances to achieve a more balanced representation, thereby improving model performance and addressing the challenges posed by imbalanced datasets.

7 WHY THESE MODELS?

7.1 ADASYN

(Haibo He et al., 2008) offers a valuable tool for improving startup success prediction. By addressing the challenges posed by imbalanced datasets, where the minority class (successful startups) is underrepresented, ADASYN enables the generation of synthetic data points for the more difficult-to-learn minority class examples. This rebalancing of the dataset reduces bias and enhances the performance of machine learning algorithms in predicting startup success. Integrating ADASYN into the prediction process involves preprocessing the dataset to create a more balanced representation of successful startups through synthetic sample generation. This approach enhances the accuracy of models and facilitates a better understanding of the factors driving startup success. Incorporating ADASYN into other prediction empowers researchers and practitioners to overcome data imbalance challenges.

7.2 Logistic Regression

(Ali and Jabeen, 2022) employed logistic regression analysis to uncover the underlying determinants influencing individuals' propensity for embarking on entrepreneurial ventures. Logistic regression, a statistical technique, was chosen to examine the association between a dependent variable, start-up intention, and various independent variables. The independent variables encompassed demographic characteristics, attitudes towards entrepreneurship, subjective norms regarding entrepreneurship, and perceived behavioural control in relation to entrepreneurship. Through the logistic regression model, the researchers calculated the likelihood of start-up intention by considering these independent variables. The regression equation was expanded to incorporate the factors influencing the inclination towards starting a business.

7.3 Decision Trees

(Wang et al., 2020) aimed to predict the fundraising outcomes of crowdfunding projects by utilizing both deep learning and commonly used machine learning algorithms. The study findings revealed that the deep learning model exhibited superior predictive performance, followed by the decision tree model. This investigation highlights the significant advantages of deep learning across various evaluation criteria, demonstrating its potential in forecasting crowdfunding project

7 WHY THESE MODELS?

7.1 ADYSN

(Haibo He et al., 2008) offers a valuable tool for improving startup success prediction. By addressing the challenges posed by imbalanced datasets, where the minority class (successful startups) is underrepresented, ADASYN enables the generation of synthetic data points for the more difficult-to-learn minority class examples. This rebalancing of the dataset reduces bias and enhances the performance of machine learning algorithms in predicting startup success. Integrating ADASYN into the prediction process involves preprocessing the dataset to create a more balanced representation of successful startups through synthetic sample generation. This approach enhances the accuracy of models and facilitates a better understanding of the factors driving startup success. Incorporating ADASYN into other prediction empowers researchers and practitioners to overcome data imbalance challenges.

7.2 Logistic Regression

(Ali and Jabeen, 2022) employed logistic regression analysis to uncover the underlying determinants influencing individuals' propensity for embarking on entrepreneurial ventures. Logistic regression, a statistical technique, was chosen to examine the association between a dependent variable, start-up intention, and various independent variables. The independent variables encompassed demographic characteristics, attitudes towards entrepreneurship, subjective norms regarding entrepreneurship, and perceived behavioural control in relation to entrepreneurship. Through the logistic regression model, the researchers calculated the likelihood of start-up intention by considering these independent variables. The regression equation was expanded to incorporate the factors influencing the inclination towards starting a business.

7.3 Decision Trees

(Wang et al., 2020) aimed to predict the fundraising outcomes of crowdfunding projects by utilizing both deep learning and commonly used machine learning algorithms. The study findings revealed that the deep learning model exhibited superior predictive performance, followed by the decision tree model. This investigation highlights the significant advantages of deep learning across various evaluation criteria, demonstrating its potential in forecasting crowdfunding project financing. The research also combined machine learning techniques with Internet finance, providing valuable insights and practical implications for future studies. To achieve their research objectives, the authors incorporated the decision tree model, a widely employed machine learning algorithm, for predicting crowdfunding fundraising outcomes. The implementation of the decision tree model was carried out using the scikit-learn library, and parameter tuning was performed using the grid search method. Cross-validation techniques were utilized to mitigate the risk of overfitting, and optimal parameters for the decision tree model were determined through the grid search process.

7.4 Random Forest

(Abhinand and Poonam, 2022) aims to employ machine learning techniques to predict the success of Indian startups. The researchers gathered data by web scraping multiple websites, compiling a dataset featuring various features describing both unicorn and non-unicorn startups in India. They utilized several machines learning classifiers, including Naïve Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, and K-Nearest neighbour, to forecast the success of startups. Furthermore, they explored the efficacy of a stacked ensemble model incorporating all the aforementioned classifiers to enhance prediction accuracy. The models underwent evaluation using metrics such as Accuracy, F1-Scores, and AUC scores. Results revealed that the Random Forest Classifier exhibited the highest accuracy of 78.8% and an AUC score of 0.79. Additionally, the stacked ensemble model achieved an accuracy of 80.1%. The researchers applied the model to predict the success rate of startups featured in the reality show, Shark Tank India, during its first season. The study's findings suggest that the proposed model can effectively predict the success of Indian startups, serving as a valuable tool for investors in making informed investment decisions.

7.5 XGBoost

(Agarwal, 2023) proposes a machine learning-based solution to predict the success of a startup. The authors use ensemble methods such as LGM Classifier, XGBoost, and AdaBoost Classifier for training and SVM for classification instead of the traditional SoftMax function. They also use margin loss instead of a standard entropy-based algorithm for the loss function and SMOTE and Tomek's links for balancing the data as the dataset is imbalanced. The approach produces 90.2% accuracy, precision, and F1 score, significantly improving various existing models. Therefore, XGBoost can be used as one of the ensemble methods for training the model to predict the success of a startup.

8 RESEARCH GAP

Embarking on a quest to address the gaps in existing research, we propose a comprehensive approach to accurately predict startup success. This approach leverages XGBoost, Random Forest, and ADASYN, and is designed to be applicable globally. We aim to use ensemble learning methods to enhance prediction accuracy, with feature selection made efficient by XGBoost's innate feature importance metric. The interpretability of these models fosters a synergistic relationship with human expertise, while their capacity for multi-class classification allows for nuanced outcome predictions. Finally, the use of ADASYN overcomes the challenge of imbalanced datasets, ensuring robust and reliable predictions. In essence, we seek to provide a reliable, globally relevant, and nuanced prediction model for startup success using advanced machine learning methodologies.

9. REQUIRED RESOURCES

Hardware and software requirements are as follow:

9.1 Hardware requirements:

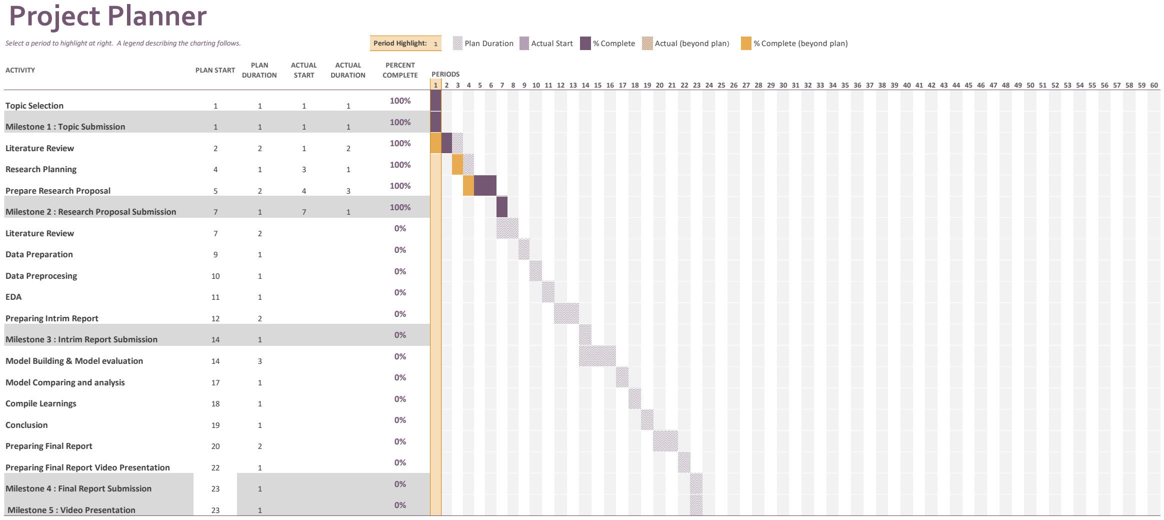
A computer with high processing power and a large storage capacity (at least 16GB of RAM, quad-core processor, and dedicated graphics card with at least 4GB of VRAM)500GB of storage space

9.2 Software requirements:

- Python programming language version 3.8 or higher for data analysis and machine learning tasks
- R programming language version 4.0 or higher for statistical analysis and data visualization
- TensorFlow version 2.4 or higher, an open-source library for machine learning and deep learning tasks
- Scikit-learn version 0.24 or higher, a machine learning library for Python
- Keras version 2.4 or higher, an open-source neural network library for Python
- Matplotlib version 3.3 or higher, a plotting library for Python
- Seaborn version 0.11 or higher, a data visualization library for Python
- Pandas version 1.2 or higher, a data analysis library for Python
- Numpy version 1.19 or higher, a numerical computing library for Python
- Jupyter Notebook version 6.1 or higher, an interactive computing environment for Python
- Data visualization and analysis tools such as Tableau version 2021.1 or higher or PowerBI version 2021.1 or higher
- A version control system like Git version 2.29 or higher to track and manage changes to the code and data throughout the research process.

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 AI for operational intelligence to improve sustainability in the beer brewing process and to develop a framework for the ethical and responsible use of AI in this context.

10. RESEARCH PLAN



Project Planner

Select a period to highlight at right. A legend describing the charting follows.

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Topic Selection	1	1	1	1	100%
Milestone 1 : Topic Submission	1	1	1	1	100%
Literature Review	2	2	1	2	100%
Research Planning	4	1	3	1	100%
Prepare Research Proposal	5	2	4	3	100%
Milestone 2 : Research Proposal Submission	7	1	7	1	100%
Literature Review	7	2			0%
Data Preparation	9	1			0%
Data Preprocesing	10	1			0%
EDA	11	1			0%
Preparing Intrim Report	12	2			0%
Milestone 3 : Intrim Report Submission	14	1			0%
Model Building & Model evaluation	14	3			0%
Model Comparing and analysis	17	1			0%
Compile Learnings	18	1			0%
Conclusion	19	1			0%
Preparing Final Report	20	2			0%
Preparing Final Report Video Presentation	22	1			0%
Milestone 4 : Final Report Submission	23	1			0%
Milestone 5 : Video Presentation	23	1			0%

11. 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.

Agarwal, K., (2023) Prediction of the Success of a Startupusing Ensemblemethods and Margin Loss.

Ali, J. and Jabeen, Z., (2022) Understanding entrepreneurial behavior for predicting start-up intention in India: Evidence from global entrepreneurship monitor (GEM) data. *Journal of Public Affairs*, [online] 221. Available at: https://onlinelibrary.wiley.com/doi/10.1002/pa.2399 [Accessed 25 Jun. 2023].

Antretter, T., Blohm, I. and Grichnik, D., (2018) Predicting Startup Survival from Digital Traces: Towards a Procedure for Early Stage Investors. [online] Available at: https://www.semanticscholar.org/paper/5ed2f23e44b9dbf656903c52f91b6a4bc4802c1e.

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.

Bangdiwala, M., Mehta, Y., Agrawal, S. and Ghane, S., (2022) Predicting Success Rate of Startups using Machine Learning Algorithms. *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, null, pp.1–6.

Böhm, M., Weking, J., Fortunat, F., Müller, S., Welpe, I. and Krcmar, H., (2017) The Business Model DNA: Towards an Approach for Predicting Business Model Success. *Wirtschaftsinformatik und Angewandte Informatik*, pp.1006–1020.

Charbuty, B. and Abdulazeez, A., (2021) Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, 201, pp.20–28.

Garkavenko, M., Gaussier, É., Mirisaee, H., Lagnier, C. and Guerraz, A., (2022) Where Do You Want To Invest? Predicting Startup Funding From Freely, Publicly Available Web Information. *ArXiv*, abs/2204.06479, p.null.