MACHINE LEARNING TO PREDICT BUSINESS SUCCESS: THEORIES, FEATURES, AND MODELS

by

Divya Gangwani

A Dissertation Submitted to the Faculty of

The College of Engineering and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Florida Atlantic University

Boca Raton, FL

December 2024

MACHINE LEARNING TO PREDICT BUSINESS SUCCESS:

THEORIES, FEATURES AND MODELS

by

Divya Gangwani

This dissertation was prepared under the direction of the candidate's dissertation advisor, Dr. Xingquan Zhu, Department of Electrical Engineering and Computer Science, and has been approved by all members of the supervisory committee. It was submitted to the faculty of the College of Engineering and Computer Science and was accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

SUPERVISORY COMMITTEE:

_____
Xingquan Zhu, Ph.D.
Dissertation Advisor

_____
Mihaela Cardei, Ph.D.

_____
Jinwoo Jang, Ph.D.

_____
Yufei Tang, Ph.D.

_____
Hari Kalva, Ph.D.
Chair, Department of Electrical Engineering
and Computer Science

_____
Stella Batalama, Ph.D.
Dean, College of Engineering &
Computer Science

_____ _____
Robert W. Stackman Jr., Ph.D. Date
Dean, Graduate College

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to everyone who has helped me throughout my journey. Your encouragement and guidance, whether big or small, have meant so much to me.

First and foremost, I am sincerely grateful to my advisor, Dr. Xingquan Zhu, for graciously accepting me into his research group and providing me with an invaluable opportunity to pursue my PhD. I extend my deepest appreciation for his unwavering support, continuous guidance, patience, and encouragement since the beginning of my PhD. Working under his mentorship has helped me expand my boundaries and pushed me to become an accomplished researcher. I am truly fortunate to have worked under him, as he has inspired me in many ways.

I would like to thank my committee members, Dr. Mihaela Cardei, Dr. Jinwoo Jang, and Dr. Yufei Tang, for their valuable insights and suggestions to help me refine my research.

I would also like to thank my lab mates and friends for all the help and suggestions they have provided me and became a source of support and inspiration in my PhD journey.

Last but not least, I am immensely grateful to my parents, Shalini and Anil Gangwani; my husband, Rohan; my siblings, Taniya and Pranav; and my extended family, Saurabh, and Suprabha, for constantly believing in me and encouraging me during the difficult time. Their support played a pivotal role in my career and personal life, making it possible to be who I am and achieve my dreams. I also express a special thanks to my baby Aryan who became the most important part of my life. Becoming a mother has not only enriched my journey but has also inspired me to complete my

PhD and serve as a role model for my son.

This research is sponsored by the National Science Foundation under grant Nos. IIS-2302786 and IIS-2236579. Any opinions, findings, conclusions, or recommendations expressed in this research are those of the author and do not necessarily reflect the views of the National Science Foundation.

ABSTRACT

Author:	Divya Gangwani
Title:	Machine Learning to Predict Business Success: Theories, Features, and Models
Institution:	Florida Atlantic University
Dissertation Advisor:	Dr. Xingquan Zhu
Degree:	Doctor of Philosophy
Year:	2024

Businesses are the driving force behind economic systems and are the lifeline of the community as they help in the prosperity and growth of the nation. Hence it is important for the business to succeed in the market. The business's success provides economic stability and sustainability that helps preserve resources for future generations. The success of a business is not only important to the owners but is also critical to the regional/domestic economic system, or even the global economy. Recent years have witnessed many new emerging businesses with tremendous success, such as Google, Apple, Facebook etc.. Yet, millions of businesses also fail or fade out within a rather short period of time. Finding patterns/factors connected to the business rise and fall remains a long-lasting question that puzzles many economists, entrepreneurs, and government officials. Recent advancements in artificial intelligence, especially machine learning, has lent researchers the powers to use data to model and predict business success. However, due to the data-driven nature of all machine learning methods, existing approaches are rather domain-driven and ad-hoc in their design and validations, particularly in the field of business prediction. The main challenge of business success prediction is twofold: (1) Identifying variables for defining business

success; (2) Feature selection and feature engineering based on three main categories Investment, Business, and Market, each of which is focused on modeling a business from a particular perspective, such as sales, management, innovation etc.

This dissertation mainly focuses on developing a framework that will aim at providing extensive features by using feature engineering and feature extraction techniques related to Investment, Business, and Market angles for forecasting business success. More specifically, the following three problems will be studied to predict success based on two different perspectives : (1) To create a triangular framework known as the IBM (Investment, Business, and Market) triangle based on important factors relevant to business success. (2) To create a modeling framework using machine learning models and graph-based models to predict the business outcome as a binary classification task. (3) To capture deep relations between different business entities using Graph-based learning algorithms and predict business success.

In summary, our major contributions to this dissertation are demonstrated in the following aspects:

- For predicting business success, identifying critical variables for defining success and selecting appropriate features is a major challenge. In order to address this issue, we first conduct an extensive study to demonstrate the most important features in different business angles and their interrelation with each other. The main goal is to predict whether the business will be successful or not. To do so, we develop additional features by carrying out statistical analysis on the dataset which highlights the importance of investments, business, and market features instead of using only the available features for modeling.

- Motivated by the above challenge, we propose a triangular framework known as IBM triangle based on three main Parities: Investment, Business, and Market. This framework provides an umbrella for defining key elements in business prediction and feature selection technique. Based on this IBM triangle, we

create a total of 563 features with three major types, including company features, investments features and market features to carry out experiments by using supervised machine learning algorithms such as Random Forest, Logistic Regression, Decision Tree, K-Nearest Neighbor, XGBoost, and AdaBoost and compare the results with the defined business target.

- To capture deeper insights into the business there is a need to establish strong relationships between different entities in the prediction task. In order to do so, the complexity of the problem increases as each of these entities has numerous attributes and interconnections between them. Hence a graph representation learning method is designed to establish four main data entities including company, investor, person, and market sector to present complex relations between these entities. Using the heterogeneous graph, a new model is proposed that provides valuable insights and a competitive edge in business prediction. The enhanced predictive analytics provides more accurate predictions when compared with other traditional machine learning models.

MACHINE LEARNING TO PREDICT BUSINESS SUCCESS: THEORIES, FEATURES, AND MODELS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Startups are the driving force of the nation due to their capability to generate economy, innovations, and employment in the country. Many companies are raising thousands of dollars and have achieved the status of a unicorn (i.e., over $1 billion) within just a few years. For example, companies like SpaceX, Canva reached unicorn status within a short time span [45]. Businesses like Uber and Microsoft are reshaping society such that new regulations have to be set in place for the upcoming businesses to keep up with these changes. Startups have created such a significant impact these days that every investor thinks about being a part of major acquisitions like Google purchasing YouTube or Facebook buying WhatsApp. However, according to the statistics, 90 percent of the startups fail during their first few years of opening. With such rapid rise and fall of businesses, it is important to keep up with the changes and factors that affect the growth of the business. The figure 1.1 shows the statistics of Startups which failed within a short span of time in spite of receiving millions of funding [32].

Our main goal is to predict the success of the business as it is not only important for the entrepreneurs or the investors but also for the nation's economy. To predict business success, we need a clear definition of what success means to the business. In our thesis, we define business success as when a company reaches two key milestones : (1) A company goes public (i.e., probability of getting an IPO), for example, when Facebook went public in 2012, allowing people to invest in the company and buy shares from the stock market. (2) A company being acquired by a larger firm or being merged with another company of the same level (i.e. Merger and Acquisition

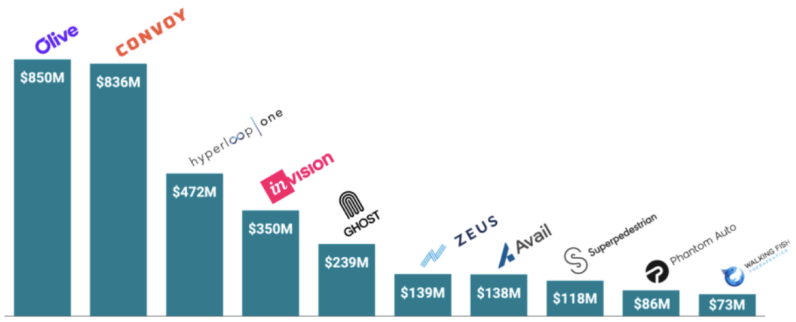


Figure 1.1: Startups failed within a short time of opening despite raising millions of funding (figure obtained from [32]).

(M&A)). For example, Google acquired Android for 50 million in 2005 [64]. This acquisition gave the company the needed tools to compete in the market. Therefore, this study will consider these two key milestones as a critical indicator for the success of the business.

With an emphasis on how these companies can benefit from improved decision-making in investment strategies and achieve financial gains when predicting business success, we use machine learning techniques and data mining to build a predictive model. The model's dependent variable is a binary label created to predict whether the company will be successful or not.

There are various machine learning applications that are applied in different fields that generate promising results. For example, in the healthcare sector, predictive modeling enables the identification of appropriate treatments for patients and even the diagnosis of illness based on patient history, images, and symptoms [60, 122]. In the finance and marketing sector, machine learning applications have been used to improve marketing strategy by creating personalized products based on customer-centric experiences, providing effective marketing strategies based on different types of customers [138, 193]. Financial sectors can also benefit from predictive modeling by managing portfolio allocation, distribution of assets, and providing personalized recommendations to customers [24, 166]. Similarly, machine learning model applica-

tions can be extended to be used to predict the success of the companies, such that the companies can make informed decisions about their investments and assets, which can lead to higher returns on their investments.

To generate a predictive model, several supervised machine learning models such as Random Forest, Logistic Regressions, SVM, and Gradient Boosting have been applied using feature engineering and feature selection processes to analyze the critical factors and create additional features that play a major role in business success prediction. We also use graph based method to predict business success by creating nodes and edges from the available dataset and identify deeper relations between them. We applied these algorithms in our experiments which demonstrated improved accuracy and AUC score when compared to the baseline for predicting business success. Successfully classifying whether the company has reached its key milestone is a crucial step for investors, entrepreneurs, and stakeholders. Moreover, the use of feature engineering and feature selection techniques to create additional features and develop models with greater predictive accuracy not only advances academic research but also has significant implications for the industry.

Although there are a lot of studies that focus on predicting business success using financial, managerial, and human resource features, most of them cannot be directly applied in practice due to the lack of knowledge about interrelated features, which is an essential requirement when predicting success. Therefore considering different features/factors that can affect the business is an important step for building a predictive model. To maximize the success of the prediction and to help the investors and stakeholders in decision-making process, it is crucial to develop a framework that considers several factors from different business angles and combines several features together to build a predictive model. A significant improvement in the results was observed using this approach when compared to the traditional methods.

Considering the improvement achieved in accuracy and AUC score with the new

framework, it is important to reinforce the advancements achieved in this study. This thesis focuses on business data analysis and explores relevant studies highlighting their importance and the objective of the study. A systematic literature review including previously researched articles related to business success, company acquisitions and mergers, and financial and marketing aspects of the business are highlighted.

1.1 OBJECTIVE

The main objective of this study is to predict whether the business will be successful or not (binary classification task). In order to do so, we build a predictive model using machine learning algorithms to classify successful vs unsuccessful companies.

Previous studies carried out using the same dataset (Crunchbase) showed that there is room for improvement as they focus mainly on financial features or business managerial features [8, 39, 88]. It is observed that by focusing only on a particular aspect of business tends to give biased results. Our study is intended to bridge the gap by providing extensive feature selection and feature engineering techniques to cover all aspects of business when classifying successful companies. Additionally, we tend to be more selective when classifying successful and unsuccessful companies based on the label created. We remove some companies due to a lack of information about whether they would be successful or not.

To fulfill our objective, our study tests different machine learning algorithms to conduct experiments on our dataset and build a predictive model.

1.2 TECHNICAL CHALLENGES

During our study, we overcame several technical challenges explained in the paragraph below.

1.2.1 Factors Relevant to Business and Business Success

In today's world, as new businesses emerge, the entrepreneurship spirit also rises due to external factors such as government funding, technical innovation, economic stability etc. such factors help people to follow their passion and encourage new business opportunities. However, as new business emerge, they also fade out within a short span of time. Hence, apart from identifying external factors, there is also a need to identify factors responsible for the rise and fall of businesses. Finding such factors that affect the business fluctuation has been a challenging task due to constant evolving of technologies, competitive market and industrial revolutions. In order to make valid predictions on business success, it is essential to consider relevant theories and factors contributing to business fluctuations. In order to overcome this challenge, we create a systematic framework based on relevant theories and studies that characterizes several factors into three main entities : Investment, Business and Market. Many established firms and startups have unique information available about their company such as their financial information, investments, funding amount, innovations, employee details etc. We characterise this available information into three main entities for prediction of business success using machine learning algorithms. However, due to the abundance of available data, obtaining insights into a business and accurately predicting its performance can be challenging, especially when considering complicated factors involved in the business operating, e.g. products, human resources, market, investment, management etc.

1.2.2 Quantifiable Features used for Learning

Another challenge we face is to identify key features in our dataset and transform them into meaningful features. Considering the important factors mentioned above and the key component required when measuring business success, we define success of the business as the capability of a company to become an IPO or be acquired or

merged with another company (M&A) of the same level. In order for a company to become successful, analyzing features and quantifying them is a necessary step in building a prediction model for business. In order to do so, we first identify available features and list them systematically based on the IBM framework. Next, we perform feature engineering and feature selection to select most appropriate features for our learning task and create additional features that are more informative and quantifiable for our prediction. For example, from the original set of features provided in the dataset, such as the "Number of investments" made by the investor within the company, we calculate the "percentage of success rate" and "percentage of failure rate" of the company. Similarly from the "first funding date" and "last funding date", we calculate "funding duration" of each company. We do this for every business angle such as for investment features, market features and business features (i.e company related features) to cover all aspects of business. however, some factors of business that provide information related to success or failure are hard to quantify; for example, innovations within the company are an important feature as they bring in new ideas and creations, which can lead to the development of new products. Hence, innovations can be measured by identifying the number of products build in the company. Therefore, these features provide all aspects of analyzing business needs that are measurable and contributes to the growth of the business. With these features, we can support different types of companies and provide a well-fitted bias-free model to predict business success using machine learning algorithms.

1.2.3 Learning Model and Performance

Business success prediction aims to analyze factors and features responsible for maintaining a profitable business and have a clear business target to produce measurable outcome. Our main challenge in the study is to develop a predictive model and conduct experiments using the dataset we have created. The goal is to produce results

that will be valuable for entrepreneurs, stakeholders, investors, and other researchers interested in furthering the research in this area. Many researchers have used machine learning algorithm for predicting the business outcome of startups or mid-size companies. Logistic Regression, SVM, and Gradient boosting have been commonly used for the learning task based on the investment features. The main aim is to develop a bias-free prediction model so VCs and stakeholders can use it in real-world prediction scenarios without hesitation. A target variable is selected based on the completion of the second round of funding and labeled as "successful" because this marks the company's stability to generate enough profit in the future. For our learning task, we apply binary classification model based on the label generated to produce desirable results. We leverage the use of a triangular relationship between investment, business, and market to develop additional features for modeling business success. The results demonstrate that adding triangular relationship-based features can indeed help improve the accuracy, compared to solutions using simple features alone. Hence, this shows that the performance of the model is effective in predicting business success and therefore fulfills our primary goal. This makes our study stand out as compared to other studies performed in this topic [7, 131, 185] due to its uniqueness in selecting quality features and creating additional features based on the proposed framework as well as using small and mid-size companies in our dataset as opposed to other studies which mostly uses only start-ups for prediction. The efficiency of the proposed method makes it reusable for further studies and even universally used for not only the entrepreneurs but also the investors and stakeholders in the company.

1.3 THESIS ORGANIZATION

Fig. 1.2 shows the thesis organization. We first study the extensive survey carried out in demonstrating and predicting different business angles responsible for defining the key elements required for business success prediction and also analyze how these

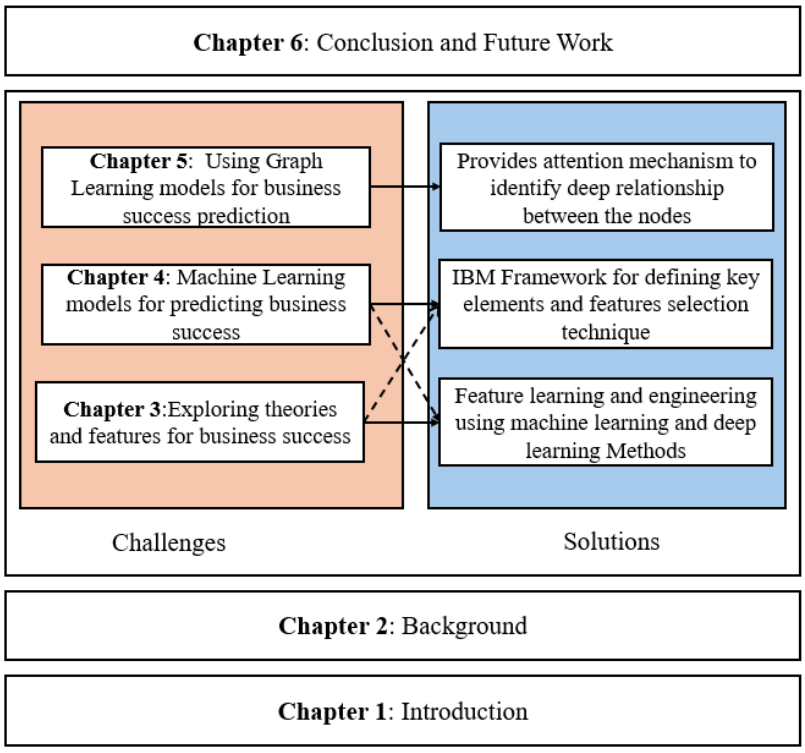


Figure 1.2: The Modeling and Predicting of Business Success. 1) Chapter 1 gives a general introduction to the studied problems with highlighted contributions and challenges. 2) Chapter 2 introduces relevant backgrounds about related work and preliminary knowledge. 3) Chapter 3 provides an extensive survey from the business perspective 4) Chapter 4 provides a new framework by exploring different angles of business, including Investor, Business, and Market, and develops a framework for predicting business success. 5) Chapter 5 studies the heterogeneous graph learning problem. 6) Finally, chapter 6 concludes the thesis and discusses future directions.

elements are interconnected to each other based on different business-related theories in Chapter 3. We then study the problem of predicting business success and exploring Investor, business, and market interplay for feature selection and feature engineering which is elaborated in Chapter 4. For this research problem, company-related data are experimentally studied, and a predictive model is built using machine learning algorithms to test the experiment results based on the proposed method. In addition to this, we also study how the company-related data can be utilized to build a graph representation learning modeling to capture rich semantics between the company, person, investor, and business sector, which is structured as a heterogeneous graph in Chapter 5.

More specifically, we organize the thesis as follows. Chapter 2 briefly describes the background including related work about modeling and predicting business success and preliminary knowledge for our proposed methods. Chapter 3 focuses on an extensive study carried out in the topic of business success prediction and demonstrates a framework that supports key factors to be taken into consideration when building a predictive model. Chapter 4 proposes a supervised learning method for predicting business success by exploring the Investment, Business, and Market angles, which are responsible for major business decisions and defining key elements for feature selection and feature engineering, which together contribute to evaluating business success. In Chapter 5, we introduce a graph-based learning method that uses heterogeneous graphs to identify relations between nodes and edges by creating four types of nodes that is company, people, market, and investors. We use an attention mechanism to identify deep relations between the nodes. Finally, in Chapter 6, we conclude the contributions and discuss the future directions of our study.

Chapter 2: Background

This chapter first provides an overview of the topic and then presents related

work about the business success prediction based on startups, mid and large size companies. This chapter also provides similarities and dissimilarities in previous studies to highlight the uniqueness of our study. Finally, preliminary knowledge are briefly described about binary classification task to predict success or failure using machine learning algorithms.

Chapter 3: Theories and Features for Business success Success Prediction

In this chapter, we first propose a comprehensive review of computational approaches for business performance modeling and prediction. For this, we first outline a triangular framework to showcase three parities connected to the business : Investment, Business and Market. After this, we align features into three main categories, each of which is focused on modeling a business from a particular point of view, for example, sales would look at the business from a marketing perspective to showcase their products and find a product-market fit, similarly management and innovation would have a different angle to look at the business. In addition, we further summarise different types of machine learning and deep learning models for business modeling and prediction.

Chapter 4: Machine Learning Models for Business Success Prediction

In this chapter, we formulate a new learning problem for business-related data and present a predictive model using machine learning algorithms for business success prediction as a solution. Specifically, we first define a new business target based on the definition of business success used in this study and then formulate additional features using the available set of features by carrying out statistical analysis on the dataset, which highlights the importance of Investment, Business, and Market angle in forecasting business success instead of using only the available features for

modeling. Finally, we apply ensemble machine learning methods as well as other supervised machine learning algorithms to predict business success and demonstrate the results using the performance metrics.

Chapter 5: Graph Learning Models for Predicting Business Success

In this chapter, we study network representation learning for a heterogeneous graph, which learns node representation features for a network using four types of nodes. We use business-related data to extract useful information and convert the dataset into a heterogeneous format to identify deep relations between different types of nodes and links in the network. A meta path-based relationship is identified between the nodes and edge type using the company data, which is used as an input to the model. A graph neural network (GCN) model is proposed, which combines the information provided as an input and predicts the outcome of the business as successful or unsuccessful.

Chapter 7: Conclusion and Future Directions

In this chapter, we summarize our contributions to Predicting Business success and their applications in the real world. We also discuss future research contributions and directions.

1.4 DISSERTATION CONTRIBUTION

In this dissertation, we summarize our major contributions to business success prediction in the following aspects below:

- To analyze the factors that are responsible for the rise and fall of the business, an extensive survey is proposed that outlines a triangular framework that demonstrates three key elements related to business: Investment, Business, and Market. These key elements are primarily responsible for business success pre-

diction and, therefore, are interrelated with each other. Using this framework, features are selected and categorized into three main categories, each of which is focused on modeling and predicting business success from a particular point of view(for example, from a sales, investor, or marketing point of view). This structured approach to business prediction not only facilitates new learning opportunities but also establishes a systematic method for forecasting success, moving beyond merely relying on existing data. Based on this framework, machine learning and deep learning methods are summarized to build a predictive model.

- In order to build a forecasting model for predicting business success, we utilize the knowledge from our extensive survey conducted on this topic and propose a triangular framework known as the IBM triangle. We explore the features diving deep into three main categories, i.e., Investment, Business, and Market, and examine their interconnection with each other. To address the challenges of business success prediction, we first define business success from a computational point of view to establish a label for our binary prediction task. Next, we perform feature engineering and feature selection processes to identify the most important features related to business success and their interrelation with each other. By combining these features from the IBM triangle, we build our dataset, which is then used for our final learning task. We apply supervised machine learning algorithms to build a prediction model and evaluate its performance using metrics such as accuracy and AUC score. Experimentation results demonstrated that using a machine learning model for binary classification tasks ensures that the model is performing well, making it suitable for further studies as well.

- In order to examine deep relations between different business entities and establish their relation in predicting business success, we propose a graph learning

network by considering business data and converting it into a heterogeneous graph structure with four types of nodes: company, investor, person, and business sector with edges connecting them. In the heterogeneous network, the presence of different types of nodes brings in a semantic relationship between nodes and edges. This complexity is generally described using meta paths that link different node types. We establish a binary classification task to predict business success using a graph neural network model using input as a heterogeneous graph with different types of nodes and edges. Using a graph learning approach for business success prediction provides a new way of looking at the businesses and also opens doors for another research angle, highlighting deeper insights about different aspects of business.

The following papers have been published & completed in relation to the dissertation study:

1. Gangwani, Divya, and Xingquan Zhu. "Modeling and prediction of business success: a survey." Artificial Intelligence Review 57.2 (2024): 44.

2. Gangwani, Divya, Xingquan Zhu, and Borko Furht. "Exploring investor-business-market interplay for business success prediction." Journal of big Data 10.1 (2023): 48.

3. Gangwani, Divya, et al. "An empirical study of deep learning frameworks for melanoma cancer detection using transfer learning and data augmentation." 2021 IEEE International Conference on Big Knowledge (ICBK). IEEE, 2021.

CHAPTER 2
BACKGROUND

This chapter first presents related work about the success of business. Then, we also present related work about factors crucial for business success and learning models used for prediction. We also describe the uniqueness in our study and compare it with previous studies in this topic.

2.1 BUSINESS SUCCESS DEFINITION AND ITS IMPORTANCE

The success of a business is a primary motivation for both investors and founders to pursue further growth and profitability. This success often translates into financial rewards that enable the company to expand its market presence. The growth of a business not only benefits investors and stakeholders but also positively impacts the nation's economy. It generates new opportunities, creates new jobs and innovations, and enhances the overall financial health of the country. This makes businesses more critical and unpredictable as it need to make changes as per the changing time. Hence, evaluating the success of a company is a critical problem that needs attention in order to maximize societal benefits [134].

Recently, many studies have focused on defining business success based on the size of companies, i.e., startup, mid-size or a large-size firm [7, 63]. The process of measuring success is different for each type of firm based on the literature studies conducted. However, most of the studies define the success of startups by focusing on two key elements,

- A company can either have an IPO (Initial Public Offering) by becoming public

and being open to buying and selling stocks of their company

- A company being acquired or merged (M&A) with another company of the same level or higher

The process of getting acquired can often lead to an exit strategy of the company, this in turn may also often mean success of the company [14, 17, 72]. For large or mid-size firms success often means external funding received by investors, and even the amount of funding received. For example, a company who achieved a status of unicorn (received valuation of $ 1 billion) is more likely to be successful and receive additional funding to continue to grow their business further [179].

Merger and Acquisition play an important aspect in corporate reorganization. It is critical for large size firms to be merged with same or higher level company as this merger provides more value to the company than being a single entity. The merger of two companies is a strategic move in order to gain more competitive advantage in the market and have more chances of succeeding in the future. Hence, M&A is an important aspect in evaluating business success. Similarly, acquisition refers to a situation where a company acquires another company with a possibility to get an exit. These companies that acquire other companies are identified as being successful.

Another aspect of defining success by many researchers is based on the survival of the firms. According to the studies, 70% startups fail during their first year of opening due to the competitive market . There are vast number of startups being opening worldwide which creates more competition and hence many startups are unable to create a product that could fit the market needs, therefore majority of the startups fail to survive. For this reason considering the survival of the firms is also an important factor when defining business success [41, 164]. Marco et al. [125], emphasizes the survival of the firms to be an important criteria when defining startup success during their initial period, particularly in comparison to companies that file for bankruptcy. However, predicting survival of the startups is extremely challenging,

as the success is dependent on external or environmental factors such as inflation, economic growth or shifts in consumer preference. Bernstein et al. [19] provided further evidence that companies that receive VC funding perform better and are likely to have a successful exit when compared to companies that do not receive funding. Such companies have higher chance of growth and bring in new innovations within the company. Few studies such as Shah et al. [157] and Krishna et al. [99] analyse business success by utilizing factors from crunchbase data and consider financial factors to determine successful companies. However, this approach includes companies that are still operational without solid indicators of future success, which could skew the modeling results and potentially misrepresent the definition of a successful business.

Considering all the aspects above for defining business success, it is evident that in our studies, we must include events that clearly define business success. For this we use two key components, that is if a company gets an IPO or is merged or acquired by another company (probably achieve an exit) of the same level to be considered as successful.

2.2 FACTORS RESPONSIBLE FOR BUSINESS SUCCESS

The success of a business is largely dependent on the company's strategic vision as well with the factors that are responsible for the growth of the business. There is a long history of researchers that study about various factors responsible for business success. Stuart et al. [168] highlighted that firms with significant market experience are more capable of developing new products and achieving success. They established a correlation between two types of variables, i.e., dependent variable responsible for initial success and independent variables such as funding, sales, market and company demographics, etc., to measure success. It was observed that by combining dependent and independent variables, one can measure success for the startups as well as mid-size companies. The study suggests that aligning entrepreneurial leadership with

technical market expertise enhances team innovation, enabling businesses to introduce new products and sustain growth effectively.

Another study by Makridakis et al [123] emphasized the importance of utilizing management tools and theories to evaluate a company's success. As market dynamics continuously evolve, businesses that stick to outdated practices are more likely to fail due to external environmental pressures. To navigate these changes, companies must adopt strategic planning, draw insights from established theories, and educate both management and staff about the market dynamics. This ensures they stay adaptable, bring in new innovations, and sustain competitiveness amid the market fluctuations which ultimately enhances their chances of long-term survival.

Several researchers have identified financial factors as crucial indicators for measuring business success [2, 103, 128, 179]. These metrics such as revenue, funding, sales, marketing, and return on investment—offer concrete data that entrepreneurs, stakeholders, and investors can leverage to make informed decisions. Industries like IT, banking, real estate, stock markets, and healthcare heavily rely on these financial indicators to assess and plan their strategies, ensuring they remain competitive and responsive to market demands. Similarly, marketing factors such as market orientation, market segment, etc., target a specific market sector for small or mid-size companies to create new products that impact the growth of the company [52, 98].

Many researchers in the past have used managerial and entrepreneurial factors for measuring success criteria in business [17,90,135]. These criterias have been proven to maximize growth and profit in small and mid-size companies. Product characteristics have also been another important factor in analyzing company's growth [90]. It is essential to have a good quality product that meets customer's needs and expectations.

Considering the important factors mentioned above and the studies carried out in the past, these factors can be grouped into three key elements that are useful in predicting business success. The three entities, namely Investment, Business, and

Market, consist of all the factors related to the business. For example, Business demographics, managerial factors that fall under Business entity, market orientation, and market dynamics are all part of a Market entity. Similarly, factors related to funding or monetary factors are part of Investment entities. There is a growing amount of literature that highlights the relationship between investment, business and market. The literature can be divided into three strands to show the relationship between these entities:

- **Investment and Business relationship:** Investments are assets or funds invested in a company with an expectation of long-term growth, specifically through M&A or becoming an IPO [134]. Ideally, the investment journey starts with planning and strategy. There is always a time frame associated with any investments which can be fruitful in long-term planning. Investments may be financial, such as foreign investment, or they may involve innovation or a company's obtaining patents. It may also be based on the business's potential or economic growth. Rai et al. [141] show the relationship between technological investments and business performance. Wan et al. [182] conducted a survey to demonstrate the relationship between foreign investment and the economic growth of the company. Many researchers show that investments in patents can have high business potential as the investors can hold or maintain their rights over the patents and gain profit [40, 55, 191]. These types of investment and business relationships show that there are high chances of success when a business finds the right paths to move forward with a plan to invest and gain maximum profit.

- **Market and Business relationship:** A business market consists of buyers and sellers who sell or exchange products and services to different consumers. Therefore, the market has a tremendous impact on business performance. Market orientation is a key factor in defining how a business reacts to the demand

in the market. Research has shown [3, 30, 67, 82, 150] that market orientation factors, such as market shifts, technological changes, product innovation and brand management, social responsibility, have a positive impact on business performance. Market resources such as *(price, advertisement, distribution)* and market knowledge capabilities [77, 129] also play a significant role in business performance.

- **Investment and Market relationship:** Investment is directly related to the market. When the market increases, such as a stock market, product sales *etc.*,, the investment also increases. Investors invest in the stock market when there is a clear growth in the company's stocks and shares [15, 181]. Based on this analysis, we can observe that investments shift to follow the market. Dot Com bubble, Blockchain applications, and Cryptocurrency exemplify how investments depend on the market [2,27,44,75]. By observing the trends and the studies, we can find that investment is always biased towards different markets.

Considering the important factors mentioned above and the studies carried out in the past demonstrated that these factors are crucial to finding the success of a business irrespective of its sector. Hence, there is a need to have a deeper understanding of these factors of a business and its organizational capabilities to evaluate business success. With the knowledge about the critical factors, we develop a systematic framework that considers these factors and creates an adaptive learning task to predict business success.

2.3 LEARNING MODELS USED FOR PREDICTION

Machine learning methods have been increasingly used in the past to predict business success [14,118]. As it is capable of leveraging vast amounts of data to identify relevant patterns in business-related data. Several researchers have used machine learning

algorithms to predict future outcomes of business-related studies like customer churn prediction [100], sales analysis, financial performance of banking sector [130], etc. These approaches include supervised algorithms like classification and regression for credit risk analysis and prediction problems and unsupervised algorithms to be applied for market trends, customer segmentation etc.

In [118] Lussier et al. used logistic regression to predict young firms' failure or success based on the data collected from survey analysis conducted on US-based firms. This survey only used a small number of businesses for its prediction with limited variables based on the available data for prediction. similar studies have been conducted on a small set of datasets that have used k-nearest neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes method for predicting successful startups.

In recent studies, researchers have used large platforms to gather huge amounts of data, unlike previous studies that used limited data for business success prediction. Krishna et al. [99] utilized the Crunchbase platform to collect data about the company's demographics and funding information such as IPO released, sales, etc, to predict business success using SVM, Random Forest, and Logistic Regression model. Using the Crunchbase platform provides vast information about the company data such as financial information, funding details, etc. [49, 192]. The data it generates contains thousands of instances to make analysis about the business. Similarly, TechCrunch is another platform that contains information about news articles about the company. Many researchers have used the TechCrunch platform to extract textual features, unlike others who use only numerical features for predicting business outcomes. Xiang et al. [186] used supervised learning methods to extract textual information from the TechCrunch platform to predict M&A in companies. The author used topic features for word extraction and then applied Bayesian Network to predict merger and acquisition.

Unsupervised learning algorithms have also been used to predict business outcome as it help in finding hidden patterns and discover meaningful information. Mainly, clustering techniques have been used in many studies along with supervised learning algorithm to predict business success [22, 62, 131]. In [22]. The authors evaluated business success using quantitative factors such as revenue generation and firms' statistical growth as key indicators. They developed a prediction model by combining k-means clustering with a Support Vector Machine(SVM) algorithm to enhance the accuracy of business success predictions. Using a combination approach, a better accuracy is achieved as compared to the previous studies.

Another aspect of the business is the Market sector, which includes market orientation, product development, customer churn rate, etc as valuable indicators of success. Managing customer churn is one of the key issues responsible for the growth of the firms. In [26], Bose et al. illustrate a hybrid model to predict customer churn behavior by employing a clustering technique to group similar customers together and then use a decision tree and boosting algorithm to predict the customer churn rate. The results demonstrated an elevated performance when including the clustering technique along with a supervised machine learning model when compared to previous studies which only focused on using supervised machine learning techniques to predict customer churn behavior.

Now a days, the studies have not just been limited to predicting company success, there has also been a large focus on studies related to business failure and bankruptcy prediction over the last few years [186]. In [156], Shah et al. demonstrated using a neural network with clustering techniques to predict bankruptcy in various firms. Three layers were used in the neural network architecture to predict bankruptcy. The first layer used financial ratio as an indicator to cluster the firms together, the second layer determined the learning process, which consisted of time series data for predicting the trend of the financial status. The third layer consisted of two neurons,

one classified bankrupt firms and the other classified non-bankrupt firms. While neural networks have shown potential in predicting business outcomes, they present challenges due to limitations in data availability and the difficulty in applying such models across diverse business contexts.

There has also been a growing amount of research using graph-based methods for predicting business success. In particular, graph theory and network analysis have been applied to demonstrate complex relationships within businesses, markets and investors. Traditional methods rely on tabular data, while graph-based methods focus on capturing deeper insights and relations between different entities like businesses, market, investors, and people. Recently a study established the use of a graph neural network of VC-invested firms by capturing rich semantics of the nodes and their neighbors to form a link between them [120]. Investments and startups were identified as nodes and the information was manually extracted for modeling graph neural networks to predict the outcome of business. This study mainly focused on the IT and Healthcare sector making it limited to be applied to other industrial sectors of the firm.

Few studies have utilized Crunchbase data, which offers comprehensive information about companies across various sectors. However, most research has relied on only a small subset of this dataset, although achieving promising results. Additionally, much of the existing work has focused on specific features, limiting the ability to capture a full picture of a company's status and potentially reducing the accuracy of business success predictions. Moreover, many of the features used in past studies may no longer be available today, making it challenging to build predictive models based on prior work. To address these gaps, our research aims to develop a systematic framework that identifies key features from previous studies and other relevant factors, then defines business success by incorporating all available company data. We highlight three critical elements namely Investor, Business and Market in

our study, demonstrating their interrelationships and collective impact on business success, which distinguishes our approach from previous studies in this field.

CHAPTER 3
THEORIES AND FEATURES FOR BUSINESS SUCCESS PREDICTION

With the rise in the global economic system in recent years, we witnessed many new emerging businesses with tremendous success such as Google, Apple,Facebook etc., yet millions of businesses also fail or fade out within a rather short span of time.

In today's global landscape, there is a notable rise in entrepreneurship, with new businesses and entrepreneurs emerging across the world. This growth is fueled by a strong entrepreneurial spirit and the pursuit of diverse business opportunities [195]. This phenomenon is driven by various factors, including government support, technological advancements, and cultural and economic influences that inspire individuals to pursue their entrepreneurial ambitions. Many entrepreneurs and private investors strive to build "unicorn" companies, such as Uber and Facebook, which have revolutionized traditional business models and significantly impacted society. However, every business venture carries inherent risks of failure alongside the potential for success. Potential risks such as not fulfilling the financial goals, poor management strategies, wrong hiring, and marketing mishaps are the most common reasons for business failures [83]. With the rise of startups and their impact on the economy, entrepreneurs, investors, and decision markers are in need of effective methods to analyze business data from different perspectives.

Identifying the key factors influencing business fluctuations remains a challenging task due to the rapid evolution of technology, increasing market competition, and ongoing industrial transformations. Recent studies have examined factors such as the likelihood of mergers and acquisitions in small or mid-size companies, financial factors

such as sales, revenue, and expenditure that lead to business success, and investments leading to IPO status [28,145]. Nevertheless, these studies have essentially limitations, because they focus on a particular method or a limited number of factors.

It is challenging to identify relevant factors and come up with an effective method to predict business success. In order to make accurate predictions, it is essential to consider relevant theories and factors that contribute to both the rise and fall of businesses. Each business follows a systematic process involving various stages or events, commonly known as the business life cycle, as depicted in 3.1. Similar to the life cycle of living organisms, every business strives to achieve success through the business life cycle. The four stages as the name suggests are key indicators of illustrating the gradual and steady growth of business and therefore play an important role in predicting business success. For example, a high growth firm can see an exponential rise in the business and hence has the capability to either decline drastically or expand their businesses through Merger and Acquisition (M&A). Based on the type of firm the life cycle of business varies in growth with respect to time. Numerous theories [124] emphasize the importance of the business cycle in identifying critical factors that minimize fluctuations and enhance the likelihood of success. Hence, understanding the growth patterns of the firms is an important factor when predicting business success [105].

Motivated by existing business studies and computer science research, our research aims to leverage theories of business fluctuations and business cycles to study essential features and factors relevant to business success. The goal is to predict the success of the business and highlight relevant methods useful to carry out the learning task. Considering the main aim of this study and the importance of business success, we summarize the following two main challenges:

- *C*hallenge 1 : How to define business success based on the criteria of success and measurable factors relevant for evaluating business success. Typically, the

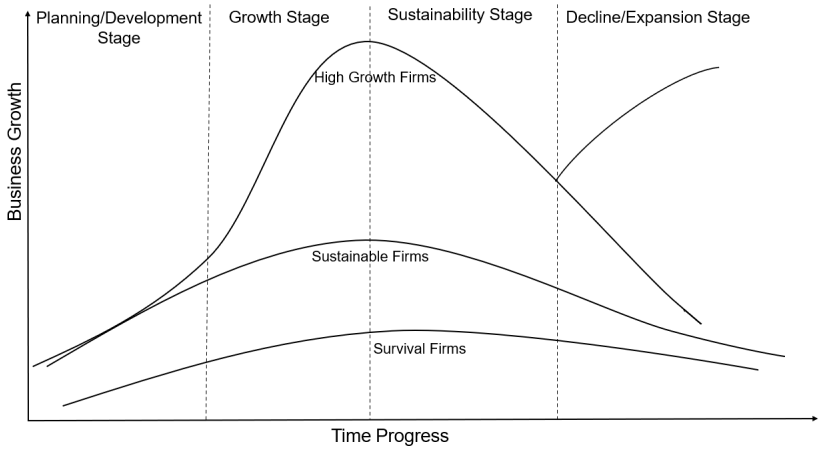


Figure 3.1: A conceptual view of five main stages of a business life cycle with the growth of three types of firms highlighting the phases and time progress

definition of success typically varies based on the type and size of the firm.

- *C*hallenge 2: How to design an end-to-end framework for predicting business success? Existing methods are either ad hoc or domain-specific, which makes them inadequate in our study.

In order to address the above challenges, we study relevant theories and literature to highlight the factors responsible for business fluctuations. For *Challenge 1*, we first define business success from a computational point of view that enables the development of a verifiable model for accurate prediction. For this, we make sure to include all types of businesses regardless of their size and type of firm. This ensures the versatility and usability of the model for entrepreneurs, investors, and stakeholders. For *Challenge 2*, we first propose a systematic framework that organizes these factors into three primary entities responsible for business success. Then, based on these three entities, we list different methods by utilizing machine learning and deep learning models for predicting business success. Our approach will benefit young entrepreneurs, investors, and even unicorn companies who are constantly seeking methods to predict business success. Our study aims to provide a comprehensive, up-

to-date literature review on business success prediction. Existing studies are focused on limited factors such as organization details, investments, and funding for a specific sector rather than including companies from all business domains. The advantage of our study is that it helps close the gap in the literature by providing a systematic framework to thoroughly review case studies, articles, and theories related to business fluctuations. A review of learning methods, data, and performance metrics further outlines the whole ecosystem of using machine learning for business success prediction. The main contribution of our study can be summarized below:

- We first study the relevant business theories and highlight case studies to provide a conceptual definition of business success.
- We propose a systematic framework known as the Investment-Business-Market (IBM) triangle framework to summarize critical factors responsible for analyzing business success. This framework serves as a general skeleton covering vital features related to the business life cycle and operation.
- We provide a detailed study of major features used for predicting business success, categorize these features according to the three main parties of the IBM triangle and explain how these features are extracted and modeled based on different business angles.
- We provide an extensive survey of machine learning models for business success prediction, which provides a landscape for researchers, investors, and entrepreneurs to understand the state of the art. It will also allow them to pivot in a timely manner so that financial resources are utilized wisely.

3.0.1 Preliminaries and Theories related to Business Success

In this section, we introduce theories of the business cycle and the IBM triangle framework to study factors and criteria for business success. We also list case studies

of successful businesses to analyze factors important for business success.

3.0.1.1 Business Cycle Theories and IBM Triangle

For centuries, the evolution of businesses, along with industrial revolutions and bankruptcies, has provided investors, entrepreneurs and researchers with valuable opportunities to analyze the driving factors behind economic changes and develop significant business cycle theories from various perspectives. A business cycle refers to the periodic ups and downs that an economy experiences, influencing different stages of business activity, either leading to growth or decline. These fluctuations or changes in the cycle, which may occur over time, can result in sudden increases in product prices or decline in the profits. Several well-established theories, such as Keynesian theory [124], have emerged and continue to shape modern business practices. The foundation behind creating the business cycle relies on economists who periodically studied the real-time GDP and investments over the years to develop a theory behind defining the four main stages of the business cycle. In the early 20th century, it was observed that businesses had seen a tremendous rise in profits and growth of the business for several years due to the technological changes in the economy, which was followed by a drastic decline that came as a shock to many businesses. The shocks such as innovations, investments, uncertainty, over-production of goods, and environmental factors affected the rise and fall of the business [33].

Various business theories are put forth by researchers to examine the key elements that influence a company's success or failure. These factors can be measurable or non-measurable such as environmental conditions. In order to protect firms from the effects of cycle variations, economists and policymakers can develop predictions based on innovations, investments, earnings, and sales by focusing on measurable criteria. Table 3.1 summarizes eight business theories and factors related to cycle fluctuations. For example, the theory of innovations [154] in business suggests that introducing new

technology and new techniques in selling a product influences continuous growth in the company and leads to long term returns on investment. While continuous innovation increases the risk of uncertainties, it enables businesses to stay competitive, ensuring a steady profit rate over time. The theory of human capital in organizations [31] highlights key competitive factors that affect the success of new companies. Factors such as business demographics and the job skills of the founders play a crucial role in determining how well an organization can endure shifts in the business cycle. A newly found organization has very little competition in the market and hence can utilize the human capital theory to capture the market's attention. A major factor that draws the attention of entrepreneurs is the market for the business. Many theories highlight marketing to be an important aspect to capture investors as well as create customer product engagement for long-term growth [96]. Allocating new marketing strategies and resources not only brings new investors into the business but also attracts customers via Business-to Business (B2B) or Business-to-Customer (B2C) relations thereby providing an advantage for the organizations to stay on top of the businesses during economical shifts. Fluctuations in investment are critical to the theory of the business cycle and must be considered when focusing on business success.

Shocks are primarily caused by the impact of investments and finances, which disrupt the patterns of the business cycle. For instance, as business capital rises in response to an increase in investments, productivity rises as well, enhancing profits and returns on capital. Conversely, a decline in investments has a significant impact on business output, which lowers profit and employment opportunities for businesses [68, 89]. Strategic planning compels an organization to adopt new perspectives in order to improve its capital expenditure, supply chain, human resources, business operations, and product pricing, according to a different study on the subject. It unfolds two major aspects which help to answer the questions 1) When should the

organization make changes and 2) how to implement those changes. When applied within the organization's workforce in a timely manner throughout the stages of the business cycle, strategic planning can minimize the effect of fluctuations.

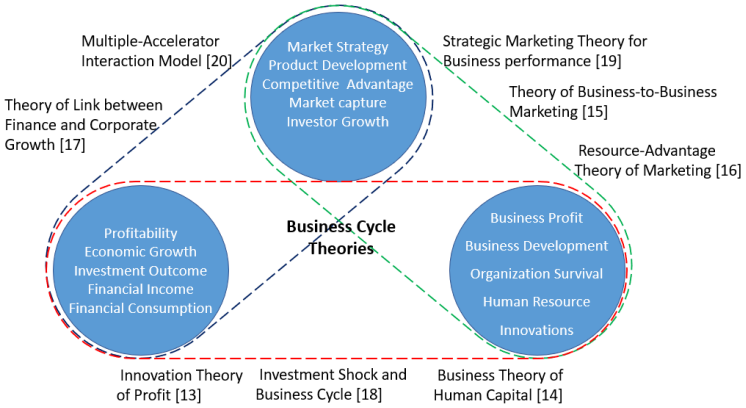


Figure 3.2: Theories of Business Cycle and the factors behind the theories

Combining theories related to major factors influencing business cycle provides a roadmap on building solutions to help businesses succeed or grow in competitive business environments. Based on business theories shown in 3.2, we propose a framework to group all factors into three main categories: investments, business, and market, which led to the creation of the IBM triangle in Fig 3.2.

Table 3.1: Summary of business cycle theories. The table lists eight theories summarizing factors/hypothesis about business evolution

Theory Name	Factors	Hypothesis
Innovation Theory of Profit (ITP) [154]	Business profit, Growth, Innovation	Creating innovation was the first step on the path to success and economic profits
Business Theory of Human Capital (BTHC) [31]	Business demographics of founders,Organizational survival,Human resource,Product development	Business theory influence the success/survival of new firms focusing on business demographics, founders details and market capture
Theory of Business-to-business Marketing (B2BM) [96]	Market strategy, Product development,Competitive advantage over b2b, Investor growth	Theory of b2b marketing focuses on attracting new customers in the market thereby capturing new investors into business
Resource-Advantage theory of Marketing (R-AM) [79]	Competitive advantage to firms, Firms success with marketing strategies, b2b and b2c competition	Resource allocation theory provides a competitive advantage over business marketing of successful firms
Strategic Marketing Theory for Business Performance (SMBP) [129]	Acquiring resources, Investments opportunity, Market capability, Business operations	Strategic market in investments, resources and organizational capabilities plays an important role in the growth of business
Theory of Link between Finance and Corporate Growth (LFCG) [69]	Financial investments, Economic growth, Net income, Profitability	The evidence from cross country studies suggest a strong link between finance and corporate growth
Investment Shocks and Business Cycle (ISBC) [89]	Stock price inflation, Investment output, Marginal wages	Investment shocks are main reason for business cycle fluctuations.
Multiple-Accelerator Interaction Model (MAIM) [149]	Private Investments, Aggregated income, Wealth distribution	Investment multiplier affect the consumption of the distribution of wealth.

3.0.1.2 The Future of Fortune 1000 Companies : Trends and Predictions

Envision a society in which giants like Google, Microsoft, Apple initiate economic expansion. Such huge enterprises fall in the list of fortune 1000 companies. They cover majority of the industrial sectors, including Technology and Healthcare. These businesses create more products and have the maximum sale. However, such business do more than just product creation and expansion. They impact the market trends, create more jobs and shape the global economy.

The success of these Fortune 1000 companies also demonstrates the usefulness in business cycle theories. For instance, Walmart topped the Fortune 1000 list of American corporations in 2022, followed by Amazon, Apple, CVS, and other companies [93]. Walmart employs about 2.3 million people and generates $572,754 million in revenue annually. The rise in hiring of employees, investing in the right product, and choosing demographic location in proximity to the product source are the main factors contributing to increased revenue growth, according to statistics.

CEOs of major corporations, including Chick-fil-A and Tifany, participated in a survey that examined the significance of measuring innovations and creating metrics that accurately capture innovation in a market that is constantly evolving. To measure how new concepts and innovations were implemented, KPIs with financial indicators were evaluated on a monthly basis [177]. It was observed that by selecting the ideal CEO for the business is the first step in building new ideas and innovations. In 2017, an article published in the Harvard Business Review (HBR) in 2017 [54] listed the top 100 leaders in the world as their company's highest-performing CEOs. When assessing and presenting the results, metrics like financial profits and non-financial indicators were utilized in evaluating and delivering results. Company's growth in terms of innovations, market capture, market expansion, proximity to the product sourcing, product delivery and Returns on Investments (ROI) was also measured. CEO's such as Pablo Isla, Jeff Bezos etc. were among the top ones capable of keeping

their businesses on the rise.

The above evidence suggests that creating a new vision with innovative ideas and generating a new business model is the key to success. Regardless of the market sector, large companies are actively using machine learning and AI to provide next-level products and services to customers. Alibaba [126] is one the leading e-commerce company utilizing Natural Language Processing (NLP) to generate product descriptions and utilizing forecasting models to predict customer-product engagement. Alphabet, a parent company of Google, relies heavily on deep learning algorithms to promote self-driving cars. Tech giants like Amazon, Microsoft, IBM, Tencent (a Chinese social media company) use machine learning, AI and cloud platforms to promote customer satisfaction, enhancing employee capabilities, product distribution, understanding customer engagement, product innovations and so on.

When taking into account the factors mentioned above, a connection can be found between attributes that are helpful for predictive models, such as the market, company demographics, product, investments, and innovations. The availability of features and elements varies based on the type of business, including large, small, and medium-sized enterprises, as shown in Table 3.2. Despite this understanding, it is still difficult to evaluate and quantify each element and examine the requirements for a successful firm.

Table 3.2: Summary of important factors to the business growth

Features/Businesses	Fortune 1000	Medium Firms	Small Firms	Startups
Business Innovation	✓	✓	✓	✓
Human Resource	✓	✓	✓	✓
Demographics	✓	✓	✓	✓
Investments	✓	✓	✓	✓
ROI	✓	-	-	-
Rate of Market Scope	✓	✓	-	✓
Product Innovation	✓	✓	✓	✓
Financial Capability	✓	✓	-	-
Market Growth	✓	-	-	-
VC Funds	✓	✓	✓	-
Technology	✓	✓	-	✓
Profit	✓	-	-	-

3.0.1.3 Investment-Business-Market Relationship for Business Modeling

Based on the previously mentioned theories and factors related to business fluctuations, Three important elements were analyzed that play an important role in predicting business success. We proposed a framework known as IBM (Investment-Business-Market) triangle as shown in Figure 3.3.

The IBM triangle is intended to serve as an umbrella framework for us to review various factors and theories related to business fluctuations. Meanwhile, it also helps answer important questions about predictive models, such as, what are the key features in predicting business success, how do these features interrelate with each other, and most importantly, what methods are available to predict business success? Depending on the data availability, IBM triangle provides a road map for feature selection and feature engineering required during model building. The interrelation

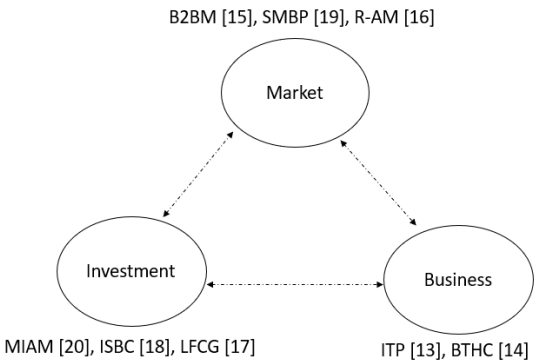


Figure 3.3: IBM triangle framework summarizing Investment, Business, and Market triangular relationship

between these three entities helps in identifying important features that can be used for model building. Hence, the factors identified previously and the correlation between the IBM entities show that our proposed IBM triangle framework is the key behind the business success prediction. As a result, the IBM triangle entities are considered to provide the basis for modeling and predicting the success of the business using machine learning algorithms. Table 3.3 summarizes business, market, and investment-related features and sub-features and their strength *vs.* weakness of using them for predictive modeling.

3.0.1.4 Business Success Criteria and Definition

There are several factors that can be used to assess how well a firm is doing. But since the majority of them are immeasurable, they can't be used directly to create prediction models. When defining success from a computational perspective, factors like the project duration, stakeholder satisfaction, staff productivity rate, etc., for instance, cannot be measured because they do not produce a clear result. We choose quantifiable metrics to assess business success based on the previously conducted research and the factors mentioned above. Table 3.4 lists important criterias for business success. Financial factors on the other hand are measurable and even

Table 3.3: Summary of features related to business success

Feature	Sub-feature	Strength	Weakness
Business Features	Demographics, administrations ,HR, education, product details, social media publicity,finances	A clear example of business features used for evaluation	Difficult to obtain complete information
Market Features	Market growth,sector, customer satisfaction, product creation, product value	Demonstrates good KPI's. Provides temporal variance	Sectors may change with time
Investments Features	Funding amount , innovations, ROI, Patents,Hiring & training cost	Most important & easily available metrics	Huge amount of redundant data

important for analyzing the business success [6, 179]. Investments and funding in the company, stock market trends and bankruptcy are always evaluated using financial criteria of the business. Marketing characteristics [52, 98] on the other hand such as market orientation, market segmentation, etc. target a specific market based on the products in small and medium size firms to provide useful factors for forecasting the business growth. Business criteria such as company demographics, business sector, managerial roles and position are proven to be useful in predicting business success. Considering important factors mentioned above and key components required when measuring business success, in this thesis, we define business success as a capability of a company to become an IPO or be acquired or merged with another company (M&A) and receive more funds from investors or VCs. On the other hand, failure is defined as a business being formally closed or bankrupted. With this definition of business success, machine learning tasks are to identify and distinguish companies between failure and success, i.e. a binary classification or multi-class classification task.

Table 3.4: A summary of business success factors and performance indicators

Business success factors	Performance indicators	Definition	References
Financial Criteria	Investments	Money invested in the company	[165]
	ROI	Return on Investment	[205]
	Funding	Received funding from VC	[168]
	IPO	Stock market value of a public company	[59]
	Seed Funding	Initial funding received	[59]
Market Characteristics	Market Growth	Potential increase in the sale of goods and services in a company	[205]
	Market Scope	The number of products and the variety of products needed based on their geographical location	[34]
	Market Need	The need of the product or services based on the location	[34]
	Market Competition	Companies selling similar products at similar price	[35]
Venture Capitalist Factors	Merger and Acquisition	Companies merging or acquiring another company for long-term growth	[34]
	Business Strategy and Plan	The number of acquisitions and strategic alliances with other company	[151]
	Financial Viability	The extent to which the company can grow or potential measure of the company's finances	[121]
	Capital Assets	The financial assets of the company	[205]
	Long-term Sustainability	Sustainability of the company over the years	[55]
	Business Partners	The number of partners the company has	[151]
	Business Valuation	Company's potential to achieve $1 billion (unicorn) valuation	[101]
Managerial Factors	Business Plan	Managers plan to execute the project within the group of employees or a team	[151]
	Team Size	The size of the team	[59]
	Technical Experience	Technical experience of the manager accessing a team	[151]
Product Characteristics	Product Quality	How good the product is compared to the similar product in the market	[55]
	Product Quantity	Ability to supply the huge amount of the product	[55]
	Uniqueness of the Product	How unique the product is	[55]
Entrepreneur criteria	Age of Entrepreneur	The age of the founder	[94]
	Skills	The skills needed for the growth of the company	[205]
	Desire for Success	The extent to which the founder is available to take risks	[174]
	Growth Ability	The ability of the founder to invest or get more funding	[121]
	Market Experience	Experience of the founders or entrepreneur in the market	[35]

3.1 THE PROPOSED METHOD

In this section we first discuss the important features related to Business success and then propose a framework to categorize important features relevant for our prediction

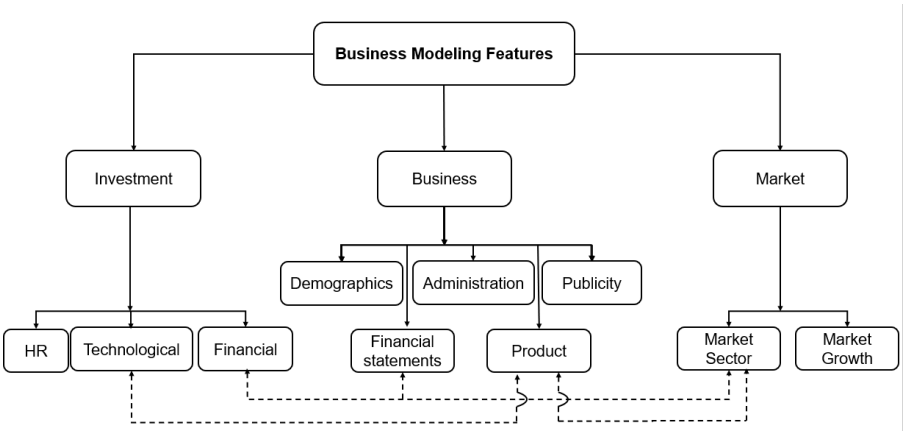


Figure 3.4: A summary of main business modeling features associated with the IBM triangle. The dashed lines show related feature subcategories.

task. Our study objective is to (1) categorize most relevant features for business modeling based on the IBM triangle and (2) present an extensive review of machine learning and deep learning models for business success prediction.

3.1.1 Features for Business Modeling

In this section, we present a detailed study of major features responsible for business success prediction. The main challenge in our study is to identify most important features and organize them as per the prediction task. For this, we propose a systematic framework known as IBM triangle. Figure 3.4 outlines primary business model features categorized under each section of our proposed framework, as well as outlines interconnection between each sub features. These feature categories list important sub-features useful for entrepreneurs to create a business strategy plan for the company.

3.1.1.1 ***Investment Features***

For a successful business, it is important to analyze investments made into the company. Not all investments need to succeed the market fluctuations. For example, the

famous Dot Com investment in *2000* was a risky move in the history of web commercialization [44] which failed and crashed the market. Therefore, before investing into the company, the investors and stakeholders nowadays evaluate all aspects of the financial features of the company to stay on top of the decision-making process. Features related to technology and human resources are also crucial components of investments. These characteristics contain vital financial data about the business, which is necessary to determine if it is profitable or not. In the below paragraph, we summarize these features in detail as it is essential for describing business in terms of funding, investments, and innovations made into the company.

- **Human Resource Features** Human Resource (HR) is one of the most valuable asset for small or large enterprises. Investments made in HR is essential to ensure the prosperity of business and changes in the market environment. Strategic investment in HR can bring a bright future to a company in terms of growth and a competitive market. Nowadays, much attention is given to the company's finances from the stakeholder's or entrepreneur's point of view to keep track of profit, loss, budget, and payroll of employees. Hence HR investments are needed in order to manage the company investments and budgeting to maintain steady growth in the market [155]. HR investments include human capital investments such as sales made by employees, cost of hiring staff, training and educating the managers, employee and team leaders, success planning, leadership development, improving work conditions of people and providing financial and health benefits to the employees. Founder-related HR features, such as job skills, titles, and roles, are also investments made by the company. Founders are the main reason companies can reach new heights by implementing new ideas and innovations in order to bring higher returns on investments [143].

- **Financial Features** During the planning stage, financial features help to effectively manage and run the business by highlighting funding and numerical goals.

When deciding whether or not to invest in a company, decision-makers must evaluate the financial aspects of the enterprise in order to identify potential risks and challenges. Perboli et al. [136] introduced a machine learning-based decision support system that predicts mid and long-term company crises for those at risk of being declared bankrupt. Financial statements of 160,000 Italian enterprises were used as an important feature to predict companies with high chances of getting bankrupt. In addition to financial features, monetary funds and financial outcomes are also responsible for predicting the success and growth of firms depending on the type and business sector. For startups, measuring them on the basis of their financial outcome is tough as they are relatively new in the market. For new ventures, measurable features such as net revenue, sales and monthly goals known as initial success by the investors are used to evaluate startup growth.

- **Technological Features** Technological features provide new advancements and innovations in the business. It attracts venture capitalist investments as technological features are responsible for economic growth. Features such as Patents, innovations or other proprietary technologies allow VC investors to control their commercial usage and provide the right to the investors to retain their benefits to the company [81]. There are several studies that use financial features such as sales, revenue generated, funding amount, etc to evaluate business performance, but technological features or strategies are not used widely as they are difficult to measure and quantify. Hence features such as Patents bring a new horizon to the market and business planning by creating new technological development or breakthroughs of a new product that has not been invented before [12]. Therefore, patents can be used as a technological measure as a company that achieves a patent can be identified as they provide unique technology information reflecting product and market development areas. Patents

are also essential for research and development collaboration (R&D) [163]. R&D employees bring new innovative projects and technological growth, which may result in a profitable return for small and mid-size enterprises (SMEs). On the other hand, the expenditure of R&D is huge for SMEs and their resources are limited to developing new products. Therefore, they tend to collaborate with large firms to obtain the right skills and necessary resources in developing new products [102].

3.1.1.2 ***Business Features***

Business features include company-related details that are important when measuring success from a nonfinancial point of view. Previous studies [1, 11, 165, 189] suggest evaluating business from both financial and nonfinancial points of view. As the company size and outcome vary depending on the goal of the company and also for whom the prediction is being made, for example, the entrepreneur, stakeholders, or investor, it is important to consider other business-related features as well. The below paragraph summarizes important business-related features that are important for the success prediction of the company.

- **Administrative Features** Administrative features encompass details about employees' work experience, salaries, education, and the technological skills they utilize in their daily tasks. These features also include supervisory information, such as the number of investors to date, the number of managers overseeing various projects, and the size of teams managed. This data offers insight into the company's work culture, its adoption of modern technologies, and its financial strength in facing external pressures. It aids entrepreneurs and investors in managing business fluctuations and preparing for external factors that could impact business growth.

- **Demographic Features** The success of a business is influenced by demo-

graphic factors, such as the location of the company's headquarters and the products or services it offers in the market [38]. The proximity of a company's headquarters to its market sector is a leading factor associated with company growth and profitability [21]. Moreover, success in different business sectors is another important factor to consider; for example, healthcare businesses take longer to establish and set up, while tech companies have a faster success rate due to their high customer engagement. E-commerce is another sector that can succeed quickly due to its higher social engagement ratio [17].

- **Product Features** To ensure success, launching a new product on the market requires meticulous planning, resource brainstorming, and innovative idea generation. Products are vital to small and mid-sized businesses because they are a new source of income and provide chances for staff members to network with clients in various market segments [57]. Though there is a chance of failure if the market is still being defined, creating a new product necessitates careful assessment of the market's capabilities and familiarity. Therefore, in order to run a successful business, it is critical to understand every aspect of product and market distribution. Product quality is an important measure that ensures uniqueness and reduces competition in the market. There should be the right amount of products with competitive prices to increase the profit margin of the sales in the company [110]. The success of the new product highly depends on the ideas and innovations used in creating a product, the technological capabilities in the business and the right type of market together contribute to the successful product and bring growth and resources into the business [55]. For companies providing services instead of products, product features may also refer to features of the services, such as functionalities of software packages or reviews of restaurants in terms of locations, food quality, and customer/staff satisfaction.

- **Financial Features** Financial aspects of the business helps to understand company's current financial position specifically whether the company is on the path of growth or not. They use financial statements for evaluation of success. These statements are basic documents that reflect a company's financial status or performance. The statements are used by many financial institutions, government agencies, and stockholders who can analyze and come up with a good idea about the future of the company's financial aspects. It gives information about the potential risks and challenges associated with investments, buying stocks from the market, granting bank loans for education, buying a property, or investing in a new business. Financial statements generally include a balance sheet, cash flow statements, income statements, and business financial ratios. These features are measurable and companies use it to evaluate the growth every quarter. Most businesses utilize a basic formula to determine financial measures including liquidity, profitability, solvency, and efficiency as shown below [117].

$$\text{Profitability} = [\text{Net income/Revenue generated}] * 100\%$$

$$\text{Liquidity} = [\text{Current asset/ Current liability}] * 100\%$$

$$\text{Solvency} = [\text{Total equity / Total asset}] * 100\%$$

$$\text{Efficiency} = [\text{Revenue generated/ Total asset}] * 100\%$$

- **Publicity Features** Businesses promote their products and services to the general public on social media sites like Facebook and Twitter [10]. For new businesses to sell their products or raise awareness of their brand among the public at a comparatively low cost, social media platforms are crucial [152]. Many investors are drawn to the information businesses produce and post on social media platforms because they are continuously looking for fresh projects with promising outcomes. The popularity of a business may be estimated using

social media elements like the number of followers, likes, comments, and reviews left by users [88]. Business firms not only look for social media platforms for the publicity of products or services in the market but also use news articles and websites to gain popularity from the public. Companies publish information about their products on websites or news articles, and the testimonials of satisfied customers are also disseminated through different channels [186]. It gives customers confidence to spend on the product and provides reviews for potential buyers. Start-up firms and established businesses like Google or Facebook heavily rely on advertising to gain popularity.

3.1.1.3 ***Market Features***

Market characteristics drive a company's success, whether it's from investors or customers. A strong market makes it simpler for a business to advance or expand in its industry. A booming market can build or break a business [97]. Start-ups thoroughly assess every aspect of the market to predict how quickly their firm will grow. The two main types of market attributes are market sector and market growth. The majority of the information regarding the company's marketing tactics and ability to plan ahead in order to turn a profit in the business is covered by these two categories. A good marketing strategy is the utmost reason that brings revenue to the company. A company can strategize its marketing techniques to sell its products and gain profit in several ways. For example, identifying new customers who might be interested in the product as well as maintaining existing customers in the business by maintaining product quality [173], using social media platforms to market new products, analyzing the product with the correct market sector to gain utmost popularity and customer satisfaction are all important means of marketing strategy. If one looks closely across the layers, all aspects of the market, business, and investments show correlation of some kind, as seen in Fig. 5. Product features in the Business area

have a strong correlation with technological features like patents or innovations. New ideas and creations brought forth by innovations result in the development of new products for the market. Advertising and marketing techniques that are tailored to the product type and market sector are necessary when developing a new product. The financial characteristics of the business, which display the debts or cash flow ratio associated with the investments or funding into the business, also significantly correspond with the financial features of the investment section. Therefore, these features provide all aspects of analyzing business needs that are measurable and contributes to the growth of the business. With these features, we can support different types of companies and provide a well-fitted bias-free model to predict business success using machine learning algorithms.

3.1.2 Methods for Business Success Prediction

In this section, we focus on describing several machine learning and deep learning models for business success prediction. One of the most important factor in predicting success of the business is having a clear definition of target in order to produce measurable results. It is also crucial to investigate business-related features in order to be able to fit them into a predictive model. We further classify machine learning models into supervised and unsupervised learning methods depending on the company's goal and business target.

3.1.2.1 Supervised Machine Learning Models for Business Success Prediction

Supervised machine learning models are commonly used for predicting business success by analyzing historical data and identifying relevant patterns that lead to a positive outcomes. It first defines a target variable of the company with inputs containing business-related features (financial ratios, company demographics, market sector and funding amount) paired with labeled outcome (binary) i.e. success or failure. Once

trained, the model can predict future outcomes based on availability of new data. The data can be split into subset of features represented by $\mathbf{x} \in \mathbb{R}^m$ (where m is the number of features, and $\mathbf{x}$ denotes a vector), the target variable is represented by $y \in \mathbb{R}$. The supervised learning model tries to predict the value of y for a given set of features $\mathbf{x}$. Supervised learning models are categorized into two main sub types: regression and classification; both are utilized in predicting and forecasting business success. Depending on the availability of label information , these models can be further divided into three distinct classes: binary classification, multi-class classification, and continuous variable prediction, as illustrated in Table 3.5.

For example, a retail industry like fashion may use supervised learning model to predict business success by analyzing factors such as sales history, marketing cost, and customer demographics. Similarly, a startup can use supervised learning model to predict success or failure by using features such as customer satisfaction rate, investments, cash flow etc.

Multi-class classification on the other hand provides a different angle for entrepreneurs and investors to evaluate the business outcome. For multi-class classification problem businesses can asses the likelihood of securing funding and being successful by examining features like founder experience, business model, and market potential with outcome variable to have more than one class such as successful, survival or failure. Many businesses do not gain profit over time, leading to the business's downfall. Various factors contribute to business failure, such as insufficient funds from the investors, poor marketing strategies, inability to compete with similar market etc. Over time, these factors become the sole reason for the company to fail in the market. Hence, this situation puts the company in danger of filing for bankruptcy. Multi-class algorithms help investors identify the risk of failure or default and give them an idea of where the company stands regarding profit.

Inthe business prediction model, measuring the growth of the business is quan-

Table 3.5: Supervised learning models and their applications in business success prediction

Class Indicator	Business Semantics	Business Target	Features	Models
Binary	Successful *vs.* Unsuccessful	Acquired or Not Acquired [190]	Business Demographics and Financial Features	LR,SVM,gradient boosting
		IPO or Not [179]	Business Demographics and Financial Features	RF, extreme gradient boosting,LR
		Second round of funding or Not funded [108]	Business Financial Features	SVM and RF
	Fail *vs.* Healthy	Survival or not survival [70]	Business Demographics and Investment Technological Features	SVM, LR, Naïve Bayes, ANN
		Successful or Fail [142] [179] [99]	Business Financial Features	Probabilistic Neural Network, SVM, gradient boosting and Logistic Regression
Multi-class	Active, Failure, bankruptcy	Failure prediction [87] [169]	Business Financial Features	Gradient boosting, Decision tree, Logistic Regression
	Risk,failure, bankrupt	Risk of bankruptcy [204]	Business Financial and Investment Features	LR, NN, decision tree
	Acquired, funding,IPO	investment decision making in Acquired or ipo company [11]	Business Financial Features	SVM,Decision Tree, GTB
	distress, Risk, failure	Financial risk in firms [43] [92] [170]	Business Financial Features	Adaboost, Decision tree,NN
	Risk,solvency,healthy	Financial stability [137]	Business Financial Features	Decision Tree,LR,SVM,RF
	Successful,Survival,Unsuccessful	Predicting survival of the company [71]	Business Demographics and Financial Features	Naïve Bayes, Random Forest, Logistic regression
	Positive,negative,neutral tweets	Predict stock market trends [139]	Business Publicity features	ANN,LR
Continuous	Forecasting continuous growth	Predicting Profit [159]	Business Financial Features	Linear regression
	Customer churn Prediction	predicting customer behaviour churn [140]	Market growth and business product features	Regression, decision tree and ANN
	Predicting growth of the company	sales prediction [37]	Business financial and product Features	Gradient Boost, Decision Tree
	High growth firms	Sales, profit and employment growth [197] [48]	Business Financial Features	Logistic Regression ANN
	Growth range(positive, negative, zero)	sales growth [190]	Business Product Features	Decision tree (CART)

tifiable. In order to effectively evaluate the business growth, a continuous evaluation process is needed to consider various factors and variables. The business's growth varies from company to company, and this is based on the defined target variable. For example, in order to evaluate success in terms of product viability, a careful mar-

ket research is needed and features like market segment, product type, sales price, demographics, etc. are evaluated. A target variable such as market trends is used as a label for a regression task. Apart from these, various outside factors such as market, business environment, and product distribution should be considered when measuring the business's success. Hence, regression models in machine learning have proven to be very useful when predicting the growth of the business.

- Support Vector Machine (SVM) : An SVM algorithm is commonly used in binary classification tasks, as it is a powerful algorithm useful for predicting data with two classes. An SVM algorithm works by finding hyperplane to separate the two classes from each other. In a recent study [108], SVM is compared with Random Forest (RF) to classify business into two groups Fail ("closed") *vs.* Not-fail ("acquired" and "operating") using features from four major groups, including region, industry, funding rounds and domain. Both SVM and RF show similar accuracy (around 88%), but their AUC values are very low with SVM being 0.51 and RF being 0.61. Because an AUC value with 0.5 implies a random classifier, this indicates that simple SVM is ineffective for business success prediction, possibly because of class imbalance, and features used in the study are less informative for classification. Similarly another study [203] used SVM for prediction bankruptcy with a small subset of data, however the results demonstrated good performance due to combining both feature selection or parameter settings of SVM model.

- Logistic Regression (LR) : Logistic regression is commonly used for business success prediction due to its simplicity and ease of interpretation in prediction task. A previous study [205] has compared LR with SVM and XGBoost to predict successful and unsuccessful firms with target variable as 0 ('operating' or 'funding series b') and 1 ('acquired' and 'IPO'). By using an exhaustive grid search as hyper-parameter tuning, LR achieved an accuracy of 86%, however,

the recall score of 0.21 was observed which improved to 0.34 when XGBoost classifier was used in comparison.

- Decision Tree (DT): Decision tree models are used for classification as well as regression tasks. The main aim of the decision tree is to predict the value of the target variable from the given dataset.The decision tree consists of two entities, nodes and the leaves. Nodes are responsible for splitting the data based on the classification problem (*e.g.* binary) and leaves decide the final outcome or the target of the business. In recent years, various financial institutions have sought simpler and more effective models for predictive tasks within the financial sector. A recent study highlighted the application of the decision tree algorithm for credit scoring, demonstrating that it is significantly simpler than previously employed complex models, which failed to deliver satisfactory results. [162].

- Naive Bayes : Naive Bayes is another classification problem based on the Bayes theorem which states that a posterior probability of an instance $\mathbf{x}$ belonging to class y, is defined by

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})} = \frac{P(x_{i1}, \cdots, x_{im}|y)P(y)}{P(\mathbf{x})} = \frac{\prod_{j=1}^{m} P(x_{ij})P(y)}{P(\mathbf{x})} \quad (3.1)$$

where $P(\mathbf{x}|y)$ is joint conditional probability of instance $\mathbf{x}$ with respect to the class y, and $P(y)$ is the prior probability of class y. According to the Naive Bayes assumption, all features are conditionally independent given the class label y, the joint conditional probability $P(\mathbf{x}|y)$ is simplified as the product of the conditional probability of all features $\prod_{j=1}^{m} P(x_{ij})$. According to the study shown in [174] when predicting successful or failed firms from the list carefully selected features by extracting uncertainty factors from the original dataset, Naive Bayes algorithm have provided better accuracy of 77 % when compared to SVM and K-Nearest Neighbor.

- Artificial Neural Network (ANN): Artificial neural networks have been widely used for the prediction of business success or failure in crowdfunding platforms. ANN uses a backpropagation algorithm for the training process. The three layers in ANN, the input layer, the hidden layer(s), and the output layer, are responsible for carrying the information from one neuron to another and generating the desired output. In the recent study based on crowdfunding project [4], predictive modeling was performed to classify successful and unsuccessful projects using ANN. Different learning rates were applied, out of which a learning rate of 0.2 with ANN gave an accuracy of 83 %, which is beneficial for the investors to provide funding for the project.

By utilizing various supervised learning algorithms for predicting business success based on small, mid-size, or large companies, entrepreneurs, stakeholders, and other decision-makers have benefited enough to make informed decisions about their businesses and minimize the risk of failure.

3.1.2.2 Unsupervised Machine Learning Model

An unsupervised machine learning model is also commonly used to analyze business-related data by finding hidden patterns or discovering meaningful groups from a given dataset. One of the most common advantages of unsupervised learning is that it doesn't rely on labeled data to provide any information. There are four broad categories of unsupervised learning. Namely, clustering, association rule mining, outlier detection, and dimensionality reduction. Table 3.6 summarizes different categories of unsupervised learning approaches and describes business implications and targets related to each category.

- **Clustering Techniques** A clustering technique identifies similar patterns, which makes it valuable for predicting business outcomes. For instance, recommender systems analyze customers with similar behaviors (such as purchase

patterns) and group them together to suggest relevant items to them [107]. Another example of customer churn prediction [178] can also be enhanced by clustering customer profiles and integrating this with classification methods. By analyzing customers with similar behaviors, products with shared characteristics, and companies with enormous growth or failure probability, we can assess broader cluster trends and predict the likelihood of business success or failure. Many clustering methods exist for business data analysis, such as k-means clustering, partition-based clustering, density-based clustering, hierarchical clustering, and model-based clustering. Among them, due to its similarity and transparency, k-means clustering is most commonly used in business domains. k-means clustering assigns n data points to k clusters, with the k value being specified beforehand. Each cluster is assigned a centroid during each iteration based on the distance of the data points to the centroid. Given a dataset with n observations $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$ which are assigned into k subsets $\mathcal{S} = \{S_1, \cdots, S_k\}$, the main objective of k-means clustering is to minimize the squared error function denoted by:

$$J(\mathcal{S}) = \arg\min_{\mathcal{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} ||\mathbf{x_j} - \mu_i||^2 \tag{3.2}$$

where μ_i denotes the centroid of cluster i, which is calculated by using the arithmetic mean of all data points in respective clustering.

- *Association Rule Mining* Association rule mining is used to examine the purchasing behaviors of customers, which helps businesses make data-driven decisions regarding product placement, pricing, and promotional strategies. By unfolding patterns in customer purchases, businesses can identify items frequently bought together, enabling the design of more effective product bundling and cross-selling strategies. This approach has been applied in areas such as product portfolio identification [86], recommendation systems [113], and detecting shifts in market trends [91], among others. In recommendation systems,

e-commerce platforms deliver personalized recommendations considering the similarities and dissimilarities of the customer's preferences. In the study [113], recommendation rules are mined for a specific customer to provide effective recommendations between customer and item rather than using a traditional co-relation-based approach. An appropriate range is specified for calculating the $[minNumRule - maxNumRule]$ rules and a scoring threshold parameter for identifying ratings (*likes and dislikes*). Based on this rating which falls under the set of these rules, collaborative and target customers are identified to provide personalized recommendations to match customers' choices with items.

Overall, association rule mining provides valuable insights into customer behavior, product portfolio, and financial analysis to help businesses make data-driven decisions about product placement, pricing, business operations, and promotional strategies, which can ultimately lead to increased sales and business success.

- *Outlier Detection* In the business world, outliers often imply significant risks or values. Many financial sectors, such as the banking industry, credit card sectors *etc.* have employed outlier detection models for the identification and prediction of irregularities in the business domain. While outliers carry multiple forms, depending on the definition, local outliers are particularly useful because they help identify samples not complying with others within a local neighborhood. Local Outlier Factor (LOF) compares the local density of the data point with the density of the neighboring data points. Previous study [36] demonstrated the use of LOF to detect inconsistencies or fraudulent activities in the banking industry to keep up with the reputation and provide customer satisfaction.

- *Dimensionality Reduction* Dimensionality reduction is a popular technique used in various areas of data analysis, including business prediction. It is used to

simplify complex datasets by reducing the number of features (i.e., dimensions) while retaining the most important information. Wu and Chong [183] proposed a two-stage ensemble approach to improve the performance of the business failure prediction using feature selection to eliminate redundant data having little or no information about the financial features. The final subset includes carefully selected financial indicators that cover information about healthy and failed firms. Three manifold learning algorithms (ISOMAP, Liner Embedding (LE), and Local Linear Embedding (LLE)) were applied to select different subsets of features and compare their performance with PCA, which enhanced the model's performance. Another study aimed to predict business bankruptcy using financial ratios [176]. PCA was used to reduce the dimensionality of the data and identify the most important financial ratios for predicting bankruptcy. The results showed that using dimensionality reduction techniques improved the accuracy of the bankruptcy prediction model. Specifically, PCA reduced the dimensionality of the data from 14 financial ratios to 5 principal components, accounting for 91% of the total variance. Hence, these studies demonstrate that using dimensionality reduction in business success prediction is useful in improving the quality of feature selection technique and thereby providing the best results for prediction tasks.

Table 3.6: Unsupervised learning models for business success prediction

Learning Objective	Business Implication	Business Target	Features	Models
Clustering	Group churn customers based on customer behaviour [178]	Churn prediction	Earning report	k-Means Clustering
	Clustering group of firms using time series analysis of financial loss [156]	Bankruptcy predicting	Financial ratios	Neural Network Clustering
	Clustering business models based on performance [22]	Success Prediction	Investors funding and revenue generated	k-means Clustering
	Clustering firms based on bankrupt or non bankrupt situations [114]	Financial Crisis Prediction	Financial Statements	k-means and EM Clustering
	Group companies based on stock prices for the investors [20]	Stock Price increase for profit on returns	Sales and market features	k-Means, EM , Hierarchical and DBSCAN
	Group companies based on growth and performance of the firms [194]	Firms sale and profit growth	Market sector and product and service price information	k-Means Clustering
	Group products & services based on consumers sentiments towards the product [73]	Business Growth	Social Media tweets and text	k-Means Clustering, sentiment analysis (NLP)
Association	Identify useful patterns in selecting customer needs [86]	customer satisfaction	Product and sales information	Association rule mining (Apriori)
	Recommend products based on market analysis [113]	Profit sales	Market needs and product information	Association rule mining (Apriori)
	Discover relations between financial data and business operations [127]	Predict bankruptcy	Financial Statements	Apriori, partition, FP-tree growth algorithms
	Discover correlations between the customers and the market [13]	Customer churn prediction	Sales and Revenue features	Aprori and FP growth
Outlier	Identify irregularities in business decisions to enhance competitive needs [167]	Business efficiency in competitive market	Product sales and market revenue	Fuzzy logic-based outlier detection
	Predict outlying behaviour of company's financial activity [36]	Predict financial crisis	Financial statements	Local outlier factor(LOF),outlier detection model
	Identify suspicious customers based on credit portfolio in companies [56]	Credit risk and financial solvency prediction	Investment reports and loan application statements	LOF, outlier detection model
Dimensionality reduction	To identify similar patterns of financial features [183]	Business Failure Prediction	Financial Statements and cash flow	PCA, Kernal-based Self-organizing map (KFSOP)
	Analyze stock market movements for future ROIs [201]	Investment decision marking based on profit returns	Earning and Sales report	PCA and ANN
	Identify similar companies and map them based on the financial aspects [112]	Financial Distress	Product sales and financial statements	Isometric feature mapping (ISOMAP)
	To recognize risky end users and filter bad credit user [161]	Credit Risk Analysis	Business demographics and profit ratio report	PCA,LDA

3.1.2.3 Deep Learning Methods for Business Success Prediction

In recent years, various studies have utilized deep learning techniques, including convolutional neural networks (CNN), long short-term memory networks (LSTM), and deep neural networks (DNN), for business success prediction. A key advantage of these methods lies in their ability to autonomously learn new features without the need for extensive domain knowledge or manually created features. Given that business success prediction often involves textual data from social media platforms, news headlines, and the finance and banking sectors, it is crucial to convert this data into vector representations before inputting them into machine learning models to enhance predictive accuracy. We examine different approaches to converting textual data into vector form using deep learning methods. To transform sentences or textual data, such as extracting positive sentiments from news articles and social media platforms, NLP techniques like Word2Vec and Doc2Vec are commonly employed. In a recent deep learning-based business failure prediction (BFP) model [25], word embedding are utilized to convert textual data into numerical form, which is then fed into the convolutional layer as input. For example, a sentence in the form of $w = [w_1, w_2, ..., w_m]$ with length m, using the word embedding to stack them in the form of a matrix represented as $N = (n_1^T, n_2^T, ..., n_m^T)^T \in \mathbb{R}^{m*l}$ where n_i is the embedded word representation of a sentence w_i. Hence, each word in a sentence can be represented as a vector $n \in \mathbb{R}^l$ where l is the number of dimensions. The primary motivation for employing deep learning methods lies in their ability to integrate textual data with numerical data, resulting in superior performance, particularly in the financial sector. This approach provides a novel perspective for researchers in the field of business prediction, as it allows for the diversification of data sources by incorporating textual information, thereby improving predictive accuracy and insight.

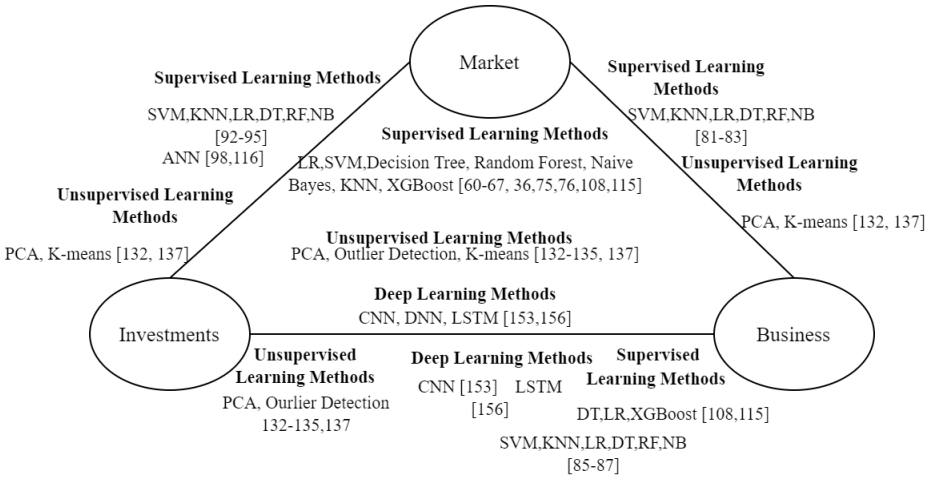


Figure 3.5: A summary of business success prediction methods and their focus with respect to the proposed IBM triangle for business modeling.

3.1.2.4 Method Summary for Business Modeling

In the above section, various methods were analyzed for predicting business success based on the size of firms and business outcomes, several methods can be applied for business modeling, such as supervised, unsupervised, and deep learning models. Each of these methods utilizes different features and outcome variables. For example, unsupervised learning methods are most commonly used for analyzing financial health and marketing products. Supervised learning methods utilize different business features to predict business success or startup survival. Recently, deep learning methods have proven to be more efficient for business failure prediction using financial and historical data. Based on the proposed framework, we highlight the methods used for business prediction in Figure 3.5. To compare the niche of three methods used for modeling and predicting business success, we summarize them in Table 3.7 concerning available features, advantages, and disadvantages.

Table 3.7: Summary of business prediction methods with IBM features

Methods	Advantages	Disadvantage	Commonly available features	May not be available
Supervised ML Models [18,34,142,152, 156,169,184]	Good control over training data	Tends to overfit high dimensional data	**Investment**-Financial features,founder info, technology **Business**-Demographics,Financial statements,Product info **Market**-Market sector/type	**Investment**-HR,Innovations **Business**-Administrative, Publicity, Web-based reviews **Market** - Market growth
Unsupervised ML Models [53,76,119,156]	Can detect hidden patterns which may not be visible to humans	Does not guarantee usefulness of results	**Investment**-Financial features **Business**-financial statements,product info **Market** -Market sector/type	**Business**-Administrative, Publicity, Web-based reviews **Market** - Market growth
Deep Learning Models [5,146,180]	Does not require fine-tuning, capability to use textual features	Requires more memory to train the model	**Investment**-Financial features **Business**-financial statements,publicity, customer review **Market** -Market sector	**Business**-Administrative features

3.2 DATASET AND RESOURCES

In this section we describe and list various publicly available dataset gathered from various platforms for business success prediction. We then review various performance metrics commonly used to evaluated modeling results.

3.2.0.1 Data Sources for Business Modeling

For business modeling and prediction, there are various datasets available on the web as well as on other platforms for researchers to extract and gather meaningful insights. The most popular and on-demand website is Crunchbase.com . The Crunchbase dataset is publicly available dataset that provides detailed information about public and private companies. It contains important information about companies, such as the investors, founders, acquisition details, funding rounds and employee details, that are needed to predict business success depends on the company's target

Table 3.8: Data and Resources

Dataset	Description	Records/Data information
Crunchbase [205] [145] [134] [111] [108] data.crunchbase.com	It is a platform that provides business information such as investors, founders, funding rounds, organization information, employee details *etc.* of public and private firms.	Structured data containing daily CSV snapshot of 50 records for categories such as (company, people,organization,funding round,acquisition).
TechCrunch [99] [186] [158] data.world/aurielle/techcrunch-startup	An online newspaper platform containing information about startups and cutting-edge technological firms available in the market. It is a platform for latest market trends and online advertisement about the companies and investors	Data contains 212 records of news articles from(company, people and products) in a csv format
Kickstarter [85] [109] github.com/sdevalapurkar/kickstarter-prediction	One of the leading crowdfunding platforms which provides information about funding, investments and creativity	Contains structured data about funded projects with over 187399 records generated in a csv file
US Patent and Trademark Office(USPTO) [184] [106] [95] [102] www.uspto.gov	An agency that issues patents to investors for their businesses and new inventions. It also issues trademark registration for the Intellectual Property(IP)	Contains structured data of publicly available patent of two types(Patent assignment data(with 8.97 million patents and Patent examination research data(PATEx) with 16.5 million patents)
AngelList [16] [199] kyang01.github.io/startup-analysis	It is one of the famous crowdfunding platforms for startups and potential investors for funding	Dataset can be scraped either by using default API or github to gather information about tweets and start-up companies.

or goal of the prediction. Many researchers have used Crunchbase dataset in their study of business success prediction due to the vast information available for them to experiment with [11, 145, 186, 192, 205] Another popular dataset for business is available in Techcrunch platform which is a popular news paper in the U.S that provides up to date information regarding new product launch, review of product usage in market, market statistics and news related to technology for startup firms [**?**, 186]. Start-ups rely on such information to achieve a specific milestone or growth in the business. The data collected from these platforms provide the basis for the prediction of success in terms of the company's M&A, IPO or financial survival. Table 3.8 summarizes datasets and provides a detailed description for each of the available datasets.

Another important platform that provides details about Patents and Intellectual Property (IP) is United States Patent and Trademark Office (USPTO). Patents and IP are important for a company to receive as they mark the technological innovations and creativity which distinguish them from other companies [95, 102]. The data collected from such platforms highlights the capacity of the company to grow and have an IPO which is an important indicator of success for the investors and the stakeholders.

A major aspect of business shifts is having information about the stock market sector. As Stock market rise and fall can affect the economy of the country and even influence corporate decisions such as job market, company expansion and investments. Many financial banks release stock market data with other financial information of the company such as profit, loss, sales and cash flow [142].

Social media platforms like Twitter and Facebook have been utilized to gather key information about companies, including their follower count, product offerings, number of tweets, and product reviews [10, 152]. These public engagement metrics raise critical questions, such as "How well-known is the company?" and "How many

customers are purchasing products from these firms?" [111, 192].

Other crowdfunding platforms such as AngelList, a U.S.-based website contemplated connecting stakeholders with investors for the purpose of funding [16]. All these resources provide a better outlook for investors to decide whether to invest in a company. It is also useful for the stakeholders to know the position of the company and anticipated growth and progress of the company.

3.2.1 Performance metrics

Assessing business performance involves analyzing the factors that contribute to its success or failure. This evaluation of business metrics is done in two ways: the performance in terms of machine learning models and the performance in terms of business interests. A confusion matrix-based performance metrics is also used to highlights the actual positives and actual negatives from the predicted values.

3.2.1.1 Confusion Matrix Based Performance Metrics

Confusion matrix is the most common metric used to evaluate the model's performance for binary as well as multi-class classification task. For business success prediction, $TruePositive$, $TtrueNegative$ and $FalsePositive$, $FalseNegative$ are used to evaluate the actual and the predicted values as shown in Eqs. (3.4) and (3.5). The confusion matrix also helps to denote the Type1 and Type 2 errors statistics which is useful to evaluate business prediction models as same metrics can be used in different models for comparison between two approaches for the same problem.

3.2.1.2 Performance Metrics for Learning models

During the learning, the evaluation of the predictive modeling is done in two steps: 1) Loss function, and 2) Performance metrics. In loss function the model is evaluated

by using the in-sample-loss-minimization function.

$$\text{argmin} \sum_{i=1}^{N} \ell\left(f\left(\mathbf{x}_i\right), y_i\right) \text{over } f(\cdot) \in F \text{ s.t. } R(f(\cdot)) \leq c \tag{3.3}$$

where $\sum_{i=1}^{N} \ell\left(f\left(\mathbf{x}_i\right), y_i\right)$ calculates the mean squared error of prediction, known as the loss function, which is to be minimized, $f\left(\mathbf{x}_i\right)$ denotes predicted values and y_i are the actual values, $f(\cdot) \in F$ is denoted as the function class of the algorithm, and $R(f(\cdot))$ is known as the complexity function which is expected to be less than a constant value $c \in \mathbb{R}$.

From a performance metrics point of view, the model's performance is evaluated based on the algorithm chosen and the type of metrics it supports.

1. **F-score (or F1-score):** F-score metric is specifically used for imbalanced datasets where one class is more dominant than others. It is calculated using both precision and recall score as shown in Eqs. (3.6) and (3.7). For binary classification task, F-score is useful to validate the model performance in terms of both classes.

2. **AUC:** Another important performance metric used for both binary and multi-class classification is Area Under the receiver operating characteristic Curve (AUC). It is one of the most frequently used metric for classification problem since it is independent from the use of false positive/negative cost.

3. **Kolmogorov-Smirnov (KS):** KS chart is a measure of degree of separation between two classes (positive and negative). Based on the business target, for example customer churn prediction or market segment analysis, KS chart is a very useful metric to predict the probability of two classes so that the business can target specific set of customers.

4. **Kappa score:** Cohen's Kappa score is used to measure the probability of agreement (p_r) between two classes on the scale of 0-1 as shown in Eqs. (3.8). It

can be used in binary as well as multi-class classification problem. For example in project evaluation where the outcome of success is divided into number of success indicators. Kappa score is more useful than accuracy when dealing with imbalanced data.

$$\text{Sensitivity/Recall } = \frac{TP}{TP + FN} \tag{3.4}$$

$$\text{Specificity (True Negative)} = \frac{TN}{TN + FP} \tag{3.5}$$

$$\text{Precision (True Positive)} = \frac{TP}{TP + FN} \tag{3.6}$$

$$\text{F1-Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.7}$$

$$\text{Kappa score} = \frac{Accuracy - p_r}{1 - p_r} \tag{3.8}$$

3.2.1.3 Performance Metrics in terms of Business Interest

To evaluate business performance and interest, in the below paragraph we highlight some common measures to be taken into consideration :

1. **Return on Equity (ROE):** ROE is a financial measure calculated by dividing a company's net income by total equity. Investors and stakeholders are more interested in knowing the returns on their investments to evaluate the company's performance and decide their next course of action. ROE provides the most easily understandable metrics without digging further into the finances and investments.

2. **Debt Ratio:** Debt ratio is the percentage of total debt to total asset ratio. In an uncertain market, the risk of investments increases the company's exposure to unexpected downturns. Hence, the Debt ratio is a good evaluation metric for the stakeholders to anticipate the shift in the market and measure where the company stands in terms of profit or loss.

3. **Stocks Buyout:** Buyout of stocks is a scenario where investors acquire the original or failed company, agreeing on a lower percentage of stocks buyout such that the failed company can exit without any debt and investors can take advantage of rectifying the market with ease. In order to measure the business loss, the buyout stock price should be closer to the market value.

4. **Liquidity Measure:** As the name suggests, liquidity measure is the amount of available cash used for business purposes within a short time. Measuring liquid cash flow is essential for knowing whether the company can withstand the market's ups and downs quickly. Having low cash flow is one of the indicators to measure business loss.

5. **Total Revenue:** Total revenue is the earnings incurred by the company after selling products/goods and services. These metrics determine whether the company reached its goal either annually or semi-annually, determining profit or loss in the market.

6. **Profit:** Total profit is calculated by dividing the income by total expense. A quarterly evaluation of these metrics helps the company to stay on top and change its goals or marketing to attain a higher number.

These metrics are useful for researchers, entrepreneurs, and new innovators to develop an optimized solution and provide thorough market research on several factors considered when developing a business model.

3.3 CONCLUSION

This study addresses the research gap surrounding business cycle fluctuations and the critical factors influencing business success while examining computational methods for predicting business outcomes. To date, only a limited number of studies have investigated the factors driving business growth and the connections between business-related theories and critical features for success prediction. Without comprehensive knowledge and facts, predicting business success remains restricted to particular use cases or industry sectors. In this chapter, we first discuss various theories related to business fluctuations and utilize these theories to identify key features and factors relevant to business success. Based on this, we propose an IBM triangle framework, which highlights three distinct dimensions of business features and illustrates their interrelations in business operations. This framework offers flexibility for researchers to expand the model by incorporating additional features and factors, accommodating various types of businesses. Building on the knowledge gathered, we review various business prediction techniques using machine learning and deep learning models, comparing popular algorithms based on their business semantics, objectives, features, and models. This review examines the applicability of different machine learning and deep learning methods in relation to specific business goals. Additionally, the study offers detailed insights into the core components of machine learning approaches for business modeling and prediction. By categorizing and summarizing key factors and features essential for business modeling, this research provides a comprehensive understanding of current methodologies and serves as a valuable resource for entrepreneurs and investors, facilitating the extension of machine learning and deep learning models across different domains.

CHAPTER 4
MACHINE LEARNING MODELS FOR BUSINESS SUCCESS PREDICTION

In a constantly evolving economy, businesses are the landscape of new research and innovations. While success of the business is defined in many ways, we stick to the previously mentioned definition of success that is if a company is in a Merger and Acquisition stage or if it received an IPO by going public. While many researchers included companies that are still in "operating" stage as successful, we do not include such companies in our study as it is uncertain of what the future of these companies will be [134,148,165]. However, in their defense, the success of the companies is based on the type of firm it is. For example, Startups are often defined based on the new innovations and products they create as startups are more technology-driven when compared to large-scale firms that rely heavily on funding from the investor [171]. This makes startups more volatile and difficult to survive longer in the competing market.

Venture Capitalist (VC) firms play a vital role in large-scale companies as they provide substantial funding to the companies to sustain the competitive market with an expectation to receive considerable returns on their investments. In addition, these firms share their network, connections, and expertise to assist the firms in their growth and profitability. VCs dedicate a reasonable amount of time and effort to making the companies grow further and achieve a profit with the intention to receive either a share of the company or returns in terms of money [135, 205]. This is the case in most scenarios; however, in a study conducted in the 2000s demonstrated that VC funds underperformed and did not meet the expectation when compared to the

S&P 500 index [74]. Hence, a business success prediction model can be of utmost use to the VC investors as it can help improve the performance of the fund distribution. Analyzing businesses that are likely to succeed has been an interesting challenge for VCs to take over. Due to all the reasons mentioned above, it has become even more important to analyze the factors of business fluctuation and identify practical solutions and methodologies to predict business success.

However, with ever changing market dynamics, previous researches need to be in sync with recent factors and literature studies to be able to modify the method based on the current situation. Hence, based on the previous theories mentioned above we analyze the factors relevant for business success prediction and provide a framework for business modeling. Our aim is to bridge the gap between previous studies by : 1) identifying most accurate definition of business success such that it brings in more clarity when selecting the most critical features from different business angles that are responsible for creating a successful business;2) Based on the definition of business success and the features available for modeling, we create additional features to demonstrate the usefulness of the selected features for modeling and predicting business success.

In recent years, several small and mid-size companies have gained attention due to their capability to capture the market by creating more jobs, launching new products, capturing emerging markets, innovations, and merging with unicorns to achieve more publicity [160]. With millions of investments made by the investors and its rapid increase in achieving unicorn status, it has become even more challenging to predict whether the business will eventually succeed or fail. Another reason for this uncertainty is the lack of historical data availability to make accurate predictions about these firms. It is difficult to keep a track of this ever changing market, hence there is a need for a method, that can track these changes and make predictions using the historical data as well.

With these dynamics, the prominence of machine learning models for predicting business success is growing rapidly. Machine learning methods have proven to be reliable for handling large volumes of data, keeping track of historical data, providing suggestions based on statistical analysis, and even identifying previously unknown samples correctly. Supervised machine learning methods such as Random Forest, SVM, Logistic Regression and Boosting algorithms like Gradient boosting and AdaBoost are mostly popularly applied for business prediction using features taken from company dataset which are publicly available on TechCrunch and Crunchbase platforms. These websites include information about company's product launch, sales and factual data [145]. In addition, many researchers also proposed neural networks in combination with classification methods to achieve high accuracy when dealing with high cardinality datasets [99]. Despite the growing amount of models built for business success prediction, most of them cannot be applied in practice due to the lack of knowledge about the interrelated features which is an essential requirement for success prediction. Moreover, many methods focus on specific features that define business success [47,88] and do not take into consideration how other features/factors can play an important role in the decisions making and in turn can result in a biased decision. In addition, many studies [17, 49, 179] gathered data from different sources, which included companies that are still in operating status and do not have enough information to determine their path toward success. Including such information may easily cause issues in trusting the applicability of the results.

Hence, In order to accurately predict business success and avoid biased results, there is a need to identify a method that includes a clear definition of business success along with useful features and interrelated sub-features that can be applied in practice to predict success. Our previous study has systematically reviewed major research challenges when predicting business success along with identifying relevant features and sub-features.

The main contribution of our work, compared to existing research in the same field, is threefold:

- Business Success : There are several types of research that define the success of a business based on the type and size of the firm [63,67]; when we define business success, we include types of the firms such as small, startups, mid-size or large-size firms. Although, as stated in the previous chapter, the overall definition of business remains the same, in this chapter, we include more factors that affect how we measure business based on different business angles. Predicting business success is intuitively important for the stakeholder and entrepreneurs who are in constant need of a method to analyze the business in a timely manner. This gives them advantage to stay on top of the market and make informed decisions about their business such as whether to invest or exit from the firm.

- Investor-Business-Market Interplay : According to the literature survey and the study mentioned in the previous chapter, there exists factors that define the success criteria for businesses. Based on these criterias, we define features for our input model. In order to predict whether the business will be successful or not, there are three main entities that are responsible for business success : Investor, Business and Market. These three entities need to be considered when analyzing the features for business success based on the critical factors examined when defining business success. There is a wide variety of research which demonstrates that each of these entities play a major role in business prediction and their interrelation helps us to identify features to be used for the prediction task.

- Feature Engineering for Business Success Prediction : In order to design a prediction framework, we use feature engineering and feature extraction method to characterize, investment, business and market angles into different types of

features to be able to fit in the prediction framework. Based on these available features, we create additional features using our IBM triangle. We designed a total of 563 features to model each company based on a binary label (1 or 0) defining success or failure. These features include information about company, investors and market. Our prediction model makes use of eight supervised learning algorithms to evaluate our modeling results.

4.1 THE PROPOSED METHOD

In this section, we define business success from the computational point of view based on the dataset used and create a label using the available features from the final dataset. Then, we perform feature engineering task to align features using our IBM triangle and make it ready for our final prediction model. Finally, we propose our business success prediction model to classify companies as successful or not.

4.1.1 Business Success and IBM Triangle Interplay

Business success is dependent of various factors such as financial condition, market scope, product development, business demographics and so on. In previous studies, we have highlighted major factors responsible for business success. Based on these studies, we concluded that there are several factors that are directly related to business success and by considering those factors we define success in terms of business. For this study, we included all types of firms, small, mid-scale, startups, unicorn and large firms when defining business success using machine learning models. In order to measure success, the outcome of the business having a status of either an IPO or Merger and Acquisition (M&A) is considered as a variable of success and companies that have outcome of closed are considered as failure. These two outcomes directly become our target variable having a binary label of 1 and 0.

As mentioned in the previous chapter, there are several factors or features identi-

fied as a measure of success. All these features, such as financial indicators, managerial features, business demographics, market sector, HR features, etc., are all interrelated and cannot be considered as separate entities for predicting business success as the dynamics of the business keep changing, but the factors that determine success remain the same. Moreover, in this thesis, we define business success for all types of firms, hence with this goal, we design an approach to divide these features into three main parities: Investment, Business, and Market as shown in Fig 4.2 and demonstrate how these features together contribute in evaluating business success. There is a wide range of publications that supports the interrelation between these three entities which shows that when determining business success or failure, these three entities must be taken into account as ignoring any one of these aspects could lead to an unsatisfactory outcome. Based on this , we define most important features for our learning model. The results suggests improved performance when using IBM entities for predicting business success. Hence, it is extremely important to consider IBM triangle features for predicting success.

4.1.1.1 Business success

In order to define business success from a computational point of view, we define success using the status of the company given in the dataset uses for experimentation. The status of the company is divided into four categories: (1) Operating; (2) Acquired; (3) IPO; and (4) Closed. The Fig. 4.1 states the statistics of the data about the companies in each status.

A company gets the status of operating if it is in the early stage of development or if it's just merely surviving in the market. Such companies do not have much information available to determine the future aspects of whether these companies will eventually succeed or will get funding to move from survival to the growth stage. A status of "IPO" and "Acquired" clearly demonstrate that such companies are

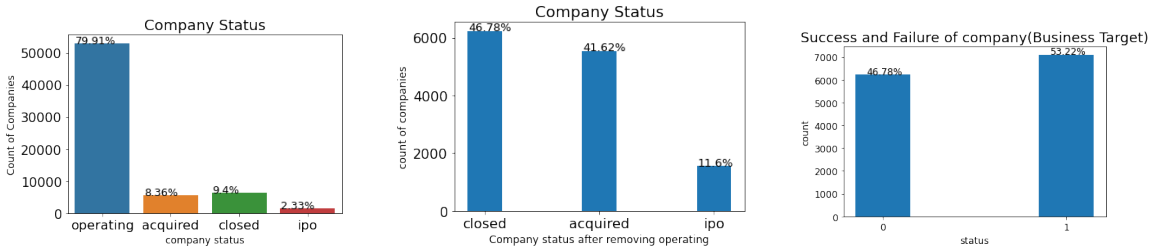


Figure 4.1: Statistics of company status with respect to four categories (a); statistics after removing Operating ones; (c) statistics of the final labels

successful firms based on the growth they achieved and the funding received from the investors. When a company goes public, it receives the status of IPO, which means that it releases its portion of its funds in the market for the public to buy a few shares of the company with the aim of achieving a huge price gain. Similarly, Merger and Acquisition (M&A) allows a company to merge with another company of the same level or get acquired by a higher level firm, for example, KPMG, EY, AT&T, etc. Therefore, when a company receives the status of IPO or Acquired, we consider them as successful for our label as it is observed that these companies have either enough funds to survive in the market for a longer duration or have just received a new round of funding to be able to sustain for few more years without any need for an external support whereas companies with the status of "closed" are no longer operating in the market is clearly considered as a failure when defining success in our dataset. Based on the company dynamics and having a clear objective of predicting success, we classified companies as successful or failure using company status as our target variable. We assigned companies with status as IPO or Acquired as 1 and companies with closed status as 0. We removed the companies with the status of operating due to a lack of information. Hence a significant portion of companies were removed from our training set as our goal is to only keep relevant information to produce significant results. Keeping such companies with little or no information would lead to biased results, as we may have to consider those companies as either

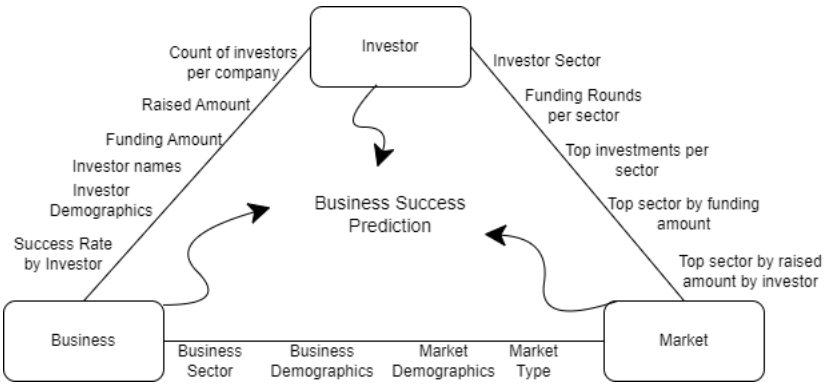


Figure 4.2: IBM triangle interplay. Investor, Business, and Market are three separate aspects that impact the business's success. Texts next to each edge outline representative features we propose to capture the interplay between them.

successful or unsuccessful. Hence, in order to have clear prediction results we removed the companies having "operating" status.

4.1.2 Feature Engineering and Statistical Analysis for Learning Task

Based on the IBM triangle framework, we define three types of features: investor features, business features, and market features. Each of these is interrelated to the other, which is useful for predicting business success using machine learning models.

4.1.2.1 Investor Features

Investor features include three main aspects of the business: Investor demographics, Investor sector, and Investor financial information. These three main features help to answer critical questions, such as which business sectors are experiencing the most growth ? how many investors have contributed to a specific market sector? and The total amount invested by each investor ? Having answers to such questions provides entrepreneurs with more information about whether the business will get repeatable returns on their investments to better assess their risk of investments into the business. Having investor information related to the business and market sector

helps in reducing the risk of uncertainty that comes with every investment made into the business.

Investor Demographics The investor demographic features include information about the investor's location, such as the city where the investor is located, country, and so on. The demographic features also include investor companies' demographic location, such as demographics of the invested company and other personal information about the investor. It is important to analyze top investors as they have more experience in identifying high-growth firms. Such investors receive enough recognition that they always have an edge toward thinking one step ahead in the game. Fig 4.3, 4.4 shows statistics of the top 10 investors in the dataset based on the amount of funding raised and the top 10 sectors in the industry that received the highest funding. This statistical analysis helps companies make informed decisions about which investor to choose in order to help their business succeed. A recent study analyzed how investor demographics directly affect the performance of the financial sector [144]. For example, a stock market industry conducted an evaluation to come up with the factors affecting the stock market's rise and fall.

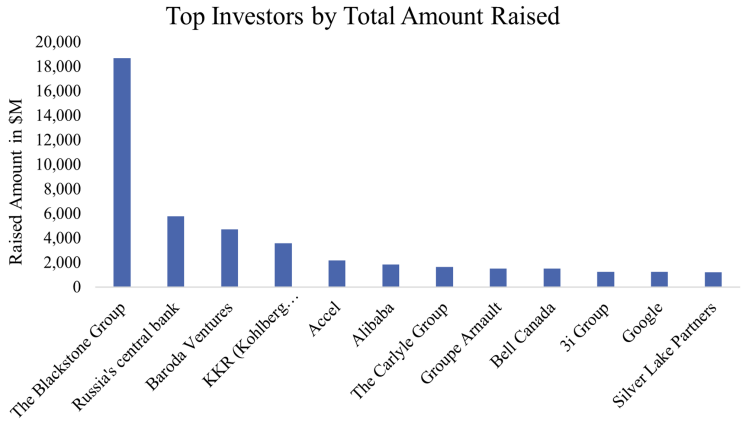


Figure 4.3: Top Investors by the Amount Raised

Investor Sector Feature Investor sector includes information about recent market trends analyzed by investors who wish to invest in a particular section. They an-

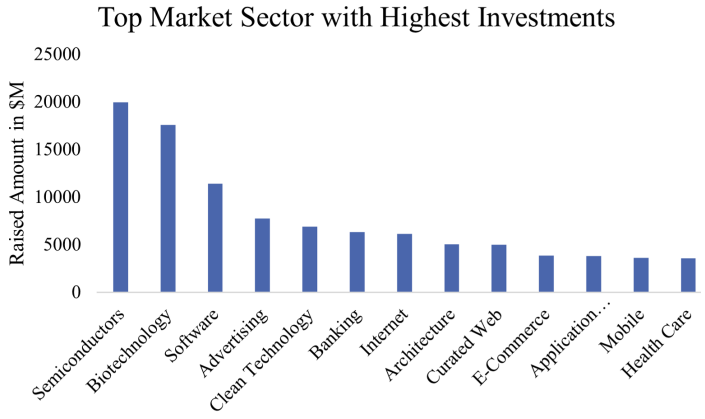


Figure 4.4: Top Market Sectors that received highest investments

alyze customer behaviour before making any decision to invest in the company. Fig shows statistical analysis on the dataset to include investor features which highlight the top market sector that received majority of the funding. As per the statistics, it is observed that Semiconductor, Biotechnology and Software are among the top 3 sectors that received maximum funding. Investing in such sectors increases the product price thereby generating maximum profit.

Investor Financial Features The financial feature includes the capability of an investor to raise more amount in the market and increase the rounds of funding such that the business and the investor achieve maximum profit. Other features such as the funding amount raised by the investor in each sector, type of funding received by the company, returns on investment *etc.,* provides enough evidence to predict the company's likelihood of success. As shown in Figure 4.5, the companies that receives the type of funding such as the Venture Capitalist (VC) funds have higher chances of being public and have faster growth rate as compared to companies backed by either seed or other types of funds [42]. Angel funding is another common type of funds that provides more chances for the company to survive the market risks and growth eventually in terms of more employment, sales and financing [104]. Investors constantly look for such new innovations and are ready to provide initial funding in exchange for a percentage of profit with these firms [196].

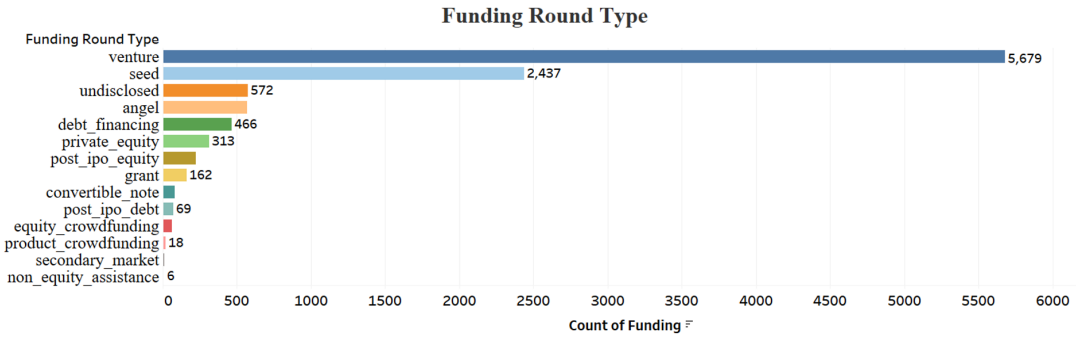


Figure 4.5: Type of funding rounds provided by investors to the companies

4.1.2.2 Business Features

Business features include company related information such as business demographics, HR features, financial information such as cash flow, amount of profit raised etc. These features are utmost important as they directly reflect company details which are important to predict business success. Many small businesses or startups do not have much information about other angles of business such as financial features, investor information such as number of investors in the company, amount raised etc [49] in order to predict success or failure. Hence, for such businesses, it is important to consider business related features as these are the only indicator of analyzing business growth. Therefore when evaluating the business success of startups or small firms, business features such as business demographics and founders' vision as well as support from the government plays an important role in giving enough information to make a successful prediction. On the other hand, for large firms, there is a need for more information such as financial features and market trends including business features to evaluate and predict business success. Regardless of the type of firms, business features serves as a common point or a major requirement when predicting business success.

Motivated by the previous studies, we identify all the important business features available in the dataset and provide a statistical analysis which gives useful insights

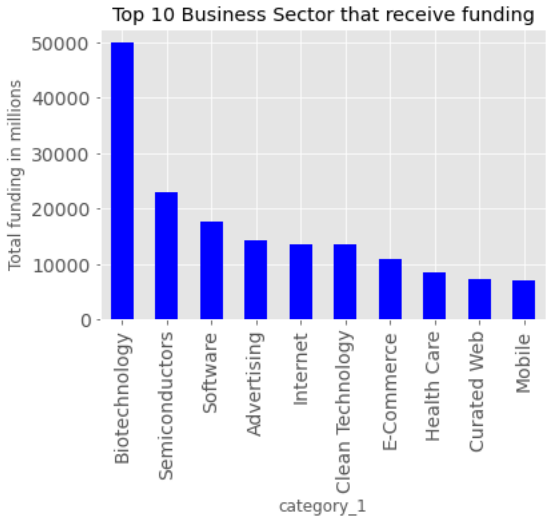
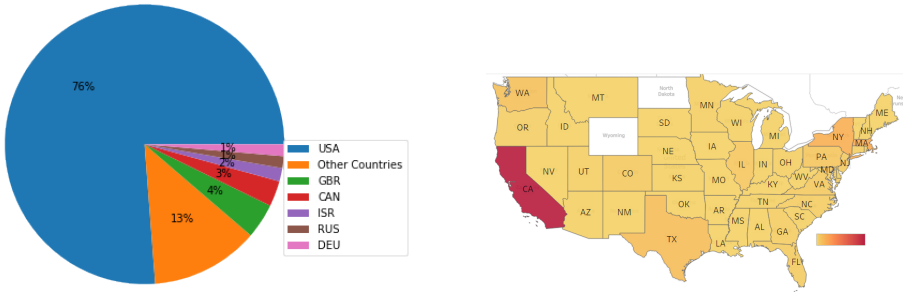


Figure 4.6: Top 10 business sectors

of the variables when developing a predictive model. For example, the analysis done on business sectors in the company demonstrates that most of the funding goes to the top business sectors that are in high demand in the market as shown in Figure 4.6. Another important factor for growth of the business is the demographics of the company that have been highlighted in Figure 4.7, which highlights that U.S has highest number of companies that are either startups or operating and within the U.S, California has the majority of headquarters locations. The sector in which the business operates is one of the important features when predicting the performance of the company as business sectors provide a sustainable environment for the companies to flourish [175]. With the ever-changing market, it is essential for the companies to keep a track of those changes and shift their investments strategies or switch funds based on the predictive methods used for analyzing market trends.

4.1.2.3 Market Features

A good market facilitates long term relationship between business and market as buyers and sellers constantly meet to exchanges goods and services. Hence market features are directly linked with business and is responsible for the rise and fall of a



(a) Percentage of companies operating across the globe

(b) Top companies withing US states

Figure 4.7: Business Demographic information

company. Features such market size, market digitization, demand and demographics determine the trends in the market and attracts customers to generate market advantage for their business.

The trends in the market constantly changes with the changing time to fulfill customers' needs. This gives the company's a reason to constantly build new products and keep the customers satisfied with constant innovations and new ideas. Figure 4.8 shows the market trends of top industrial sectors based on the revenue generated in the 2022 quarter results. As shown in the figure, Technology has generated the maximum revenue followed by Retail, Finance sector, and so on [46]. To evaluate market trends, investors gather data on revenue generated, adoption of latest technologies, level of innovation, and stock market performance over time. These key elements equip investors with the necessary insights to make informed investment decisions. Figure 4.9 illustrates the percentage of market capitalization across different sectors from 1900 to 2018 [51], highlighting shifts in market dynamics. For instance, the Transportation sector dominated in the early 1900s but saw a sharp decline by the 2000s, whereas Information Technology experienced a surge in the late 1900s and has continued to grow. This analysis underscores the importance of staying informed on market trends and other key factors that can significantly impact the success of a

business.

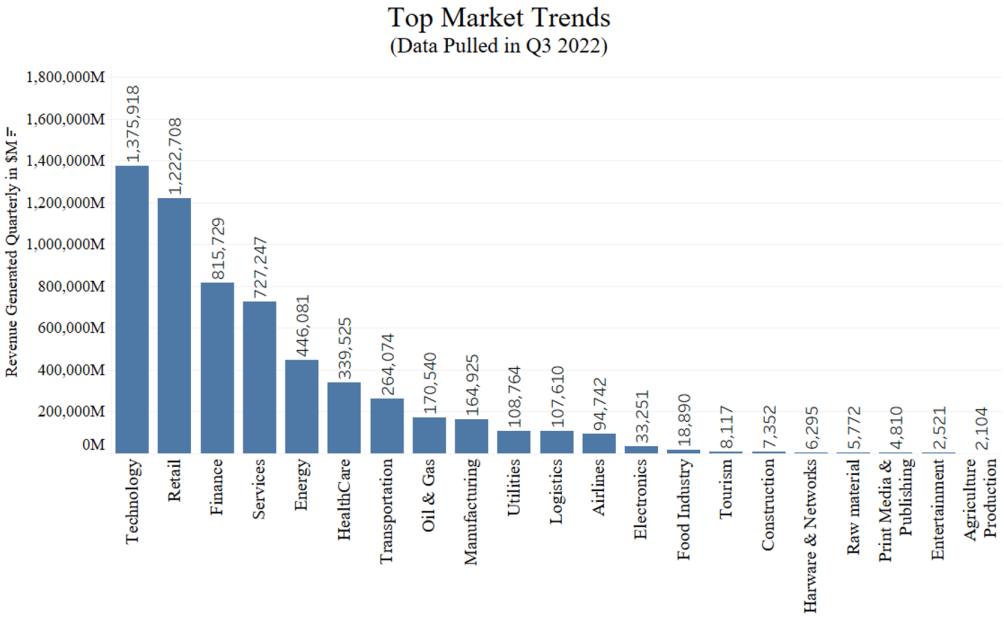


Figure 4.8: Market trend in 2022 quarter with revenue generated in millions by each sector (the plot only lists popular sectors)

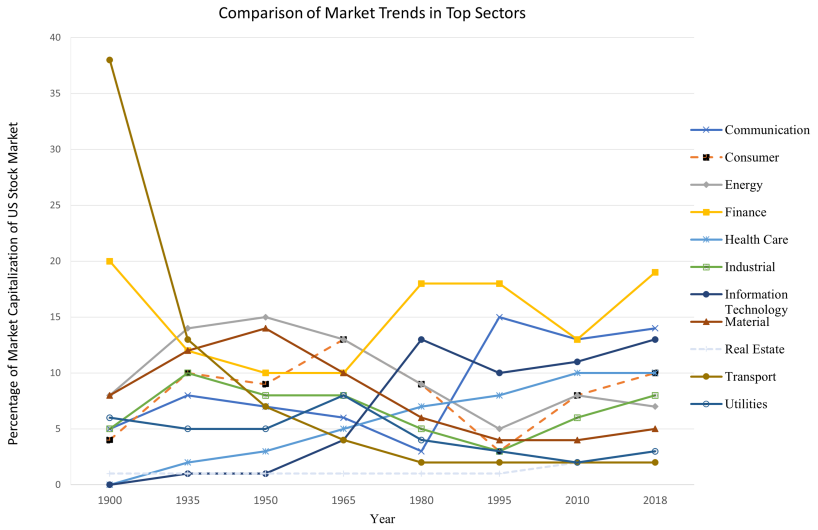


Figure 4.9: Comparison of stock market capitalization from 1900 to 2018. The y-axis shows the total percentage of market valuation of top sectors

4.1.3 Business success Prediction Model

In this section, we define the features used for learning from our dataset and then we describe the prediction framework for modeling business success.

4.1.3.1 Features for Learning

The features extracted from the dataset for machine learning task is the main step in the proposed framework for analyzing business success. The dataset taken for the learning task include two major files, the company and the investor file. These two files contain detailed information about the company, their demographics, financial aspects such as funding amount, year, etc; market sector and investor details such as investor demographics, investment amount etc. These two files are exported and merged with a unique identifier known as *Permalink* to extract meaningful features and make them ready for modeling. Three types of features are extracted from the dataset which includes investor, business, and market features that describes the correlation between IBM entities in our prediction model.

Investor Features

Investor features includes detailed information about the investors such the the name of the investor, country, state and other demographics along with features that are related to the business. In our model, we select relevant investor features including *Investor names, Funding Rounds,Types of Funding* and *Raised Amount.* The selection of these features from the investor is due to the importance of these features in business prediction and since these features are directly available from the dataset used for study [9]. Based on the available set of features, statistical analysis was performed to create new calculated set of features that provide more details about how the investments impact the company's performance. Some available features such

as the *Investor names* were one hot encoded into 14 dimensions by extracting the most common investor names (such as Angel, Venture, Bank, Technology *etc.* from the dataset). The investor names were extracted by counting the occurrences of each word and the top investor names were selected for feature extraction method. We included investor names as one of the feature as the the count of number of times each investor invested and in which sector provides details about famous investors and their business sector of operation. Other available features such as *Raised Amount* was split into 4 dimensions and scaled into *USD, millions, billions and thousands* for ease of use. Similarly, *Funding round type* and *Funding round code* provides the type of funding received by the company and the the funding code is a unique code generated for different types of funding. Using the investor names, funding information and the amount raised by investors brings in more innovations, next generation ideas and expansions into the business.

Based on the existing features and their critical importance to success, we calculated additional features to enhance the performance of the model and extract more information from the available dataset. Features such as *Percentage of success rate* and *Percentage of failure rate* is calculated by using the percentage of total investments made by the investor within the company. The business target (1 or 0) is used to distinguish success and failure by the investors. Similarly, we calculated the *Sum of Success Amount Raised* and *Sum of Failure Amount Raised* by the investors by adding the total amount raised and distinguishing it by the business target. Other calculated features are straightforward and includes *Number of investors*, *Number of successful companies*, *Number of failed companies*, and *Funded date* which is split into 3 dimensions. The *Total number of investors* is the count of investors who invested in each company. The final set of features including the description used in the modeling process is provided in the Table 4.1. The type of features are defined as either categorical or numerical and the dimension size of each feature is given after

the feature extraction and encoding process.

Table 4.1: A description of Investor Features used in the prediction model

Feature Names	Description	Type of Feature	Dimension Size after Encoding
Investor Name	Name of the investor in the company	Categorical	one hot encoded- 14 dimensions
Raised Amount $USD	Amount raised by investors into the company in USD	Numerical	1-D
Raised Amount $m	Amount raised by investors into the company in millions	Numerical	1-D
Raised Amount $b	Amount raised by investors into the company in billions	Numerical	1-D
Raised Amount $k	Amount raised by investors into the company in thousands	Numerical	1-D
Number of investors	Total No. of investors in the company	Numerical	1-D
Funding Round Type	Type of funding received by the company (seed, angel, VC ...)	Categorical	one hot encoded - 13 dimensions
Funding Round Code	Funding Codes defines the code of the funding received by the company (A,B,C ...)	Categorical	one hot encoded - 6 dimensions
Percentage of success rate by company	Success rate calculated by total number of companies invested by the investor(using business target(1))	Numerical	1-D
Percentage of failure rate	Failure rate calculated by total number of companies invested by the investor(using business target(0))	Numerical	1-D
Sum of successful raised amount	Total sum of amount raised by investors based on success	Numerical	1-D
Sum of failed raised amount	Total sum of amount raised by investors based on failure	Numerical	1-D
Total Raised Amount	Total sum of amount raised by investors including failed and successful companies	Numerical	1-D
Average funding received	Calculated the average amount of funding received by each investor to the company	Numerical	1-D
No. of successful companies	Count of successful companies by each investor	Numerical	1-D
No. of failed companies	Count of failed companies by each investor	Numerical	1-D
Funded at year	Year at which the company received it's funding by investor	Numerical	1-D
Funded at month	Month at which the company received it's funding by investor	Numerical	1-D
Funded at day	Day at which the company received it's funding by investor	Numerical	1-D

Business Features

Business features are the most crucial and backbone of our prediction framework. It provides detailed information about the companies including demographics, market sector, financial aspect and funding information. If we dive deep into the business features, the demographics include *Company name, State code, Region, Country* and *Homepage URL*. This information is useful in analyzing questions such as which region has the most startups and what is the amount of funding received by these companies. Such information provides a competitive edge to the entrepreneurs to keep up with sales and profit of the company [158]. Headquarter location of the firm attracts customers and young professionals to generate fresh ideas and market products within the proximity. Choosing the right and upcoming region as the headquarter location for new business is a great move to set their place in the market. *The Business sector*

contains the information about the market sector in which the company operates and is a common feature that is useful in defining business information as well as market characteristics. This feature is also important for the investor as they constantly look for a particular sector to invest their money in. In order to determine whether the company is growing and making profit, is to look at how many sectors the company expands their business into. Although just by evaluating market sector for success is not useful as some sectors might do well while other may fail, nevertheless, it gives an overall idea about the company's performance.

In our dataset, the sector information is present as a list of categories for each company. To extract the information, we use Count Vectorizer technique as a feature extraction method to tokenize and count the number of times the word occurs in the dataset after excluding the stop words from the dictionary. This technique maintains the semantic meaning of the words after transforming the sentence into tokens. Figure 4.10 demonstrates the example of final feature set after extraction.

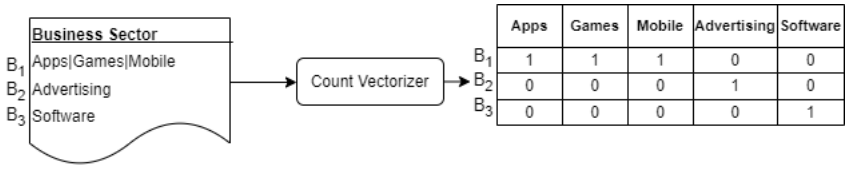


Figure 4.10: Examples of creating business sector feature using Count Vectorizer technique. Each business B_i has a list of "Business Sector" tags in the dataset (left). The Count Vectorizer represents each business B_i as one-hot (0/1) features, depending on whether a business has a specific "Business Sector" tag or not (right)

Another important information included in the business feature table is the funding information about the company. This information provides details about the initial funding received by the company based on different business sectors. It has been observed that the companies that receive initial funding or Venture capitalist funding usually perform better than the other companies that do not receive external

funding [50]. The available funding features includes *Funding Amount, First Funding Date* and *Last Funding Date*. The *Founded at* date provides the date at which the company was founded. All these features are important to analyze the situation of the company in terms of failure and success and the date features helps to analyze the growth rate of the company. Utilizing the available list of features, we have calculated additional features to support our prediction model and highlight critical aspects of the company.

With the help of the date features, we calculated *Funding duration* and scaled into days, months and years. Similarly, *Age* of the company is calculated using the *Founded at* date. The *Average duration of Funding* is calculated by finding out the average between the *first funding date* and *last funding date* by the company. Apart from this, we extracted domain information from the company features to distinguish the domain knowledge (such as .com,.net,.uk *etc*). The domain knowledge gained much popularity in the late 1990's when new era of internet grew across the world [147]. This led to the new rise in the businesses and many new companies were founded during the 1990's including the .com companies [115]. Hence, the *domain name* provides awareness among the users as it sets an established name within the company which helps to develop a certain amount of trust with the consumers and entrepreneurs. The final list of features for modeling is provided in the Table 4.2 including the description, type and dimensionality of the features after encoding. The Date features were split into month, day and year for ease of access. The following features were removed from the final dataset after analyzing the amount of incorrect information and missing values ratio within the features : *State code, region, city, company name, homepage URL and Permalink*(a unique identifier, not required after merging two dataset).

Market Features

Table 4.2: A description of Business Features used in the prediction model

Feature Names	**Description**	**Type of Feature**	**Dimension Size after Encoding**
Company domain	Domain of the company (.com,.net,.uk *etc.*)	Categorical	one hot encoded 5 dimension after encoding
Business Sector	Type of business sector in which the company operates	Categorical	Count Vectorizer - 480 dimension (taken from market features)
Company Status	Status of the company(closed, IPO, operating,Acquired) which later becomes the business target	Categorical	Binary label-1 dimension
Country_code	Country in which the company operates	Categorical	one hot encoded - 7 dimensions
Funding Total $USD	Total amount of funding initially present in the company in USD	Numerical	1-D
Funding Total $m	Total amount of funding initially present in the company in millions	Numerical	1-D
Funding Total $b	Total amount of funding initially present in the company in billions	Numerical	1-D
Funding Total $k	Total amount of funding initially present in the company in thousands	Numerical	1-D
Age of company	Age of company calculated using founded at date	Numerical	1-D
Average duration of funding	Average funding received by the company	Numerical	1-D
Funding Duration Days	The duration of the funding received by the company in days	Numerical	1-D
Funding Duration Months	The duration of the funding received by the company in months	Numerical	1-D
Funding Duration Years	The duration of the funding received by the company in years	Numerical	1-D
Founded at Day	The day at which the company was founded	Numerical	1-D
Founded at Month	The month at which the company was founded	Numerical	1-D
Founded at Year	The year at which the company was founded	Numerical	1-D
First Funding Day	Day of the first funding received by the company	Numerical	1-D
First Funding Month	Month of the first funding received by the company	Numerical	1-D
First Funding Year	Year of the first funding received by the company	Numerical	1-D
Last Funding Day	Day of the last funding received by the company	Numerical	1-D
Last Funding Month	Month of the last funding received by the company	Numerical	1-D
Last Funding Year	Year of the last funding received by the company	Numerical	1-D
Funding Round	Total rounds of funding received by the company	Numerical	label encoded - 7 dimensions

The Market features are an important link between business and investors as both look into new market trends and benefit out of it. Entrepreneurs must conduct a thorough market research before launching any new product and think about the pros and cons of current market situation during new product launch [187]. This helps them establish a customer base in different business sectors such that the investors can invest into the business freely and gain maximum profit from the product sales and services [78].

In our dataset, market feature is available in the form of business sectors which includes information about different markets sectors in which the company operates. The feature *Business sector* is included in the business feature table as well (as shown in Figure 4.10). The *business sector* is a common feature used in the market as well as business to extract useful information about the market trends, market capabilities, funding capacity for investors and the amount received per sector within the company.

In order to characterize the evolving of the market with respect to different period of time, we are creating three market features, "Top past sector", "Top current sector", and "Top future sector" to outline common business sectors, with respect to the founded time of each business. The motivation is to capture whether a business, when established, is falling into some hot market trends. An example of creating such features is shown in Figure 4.11. More specifically, *Top past sector*, *Top future sector* and *Top current sector* were created using *founded at date* of the company from the company feature table to distinguish past, current and future categories of the market. The calculation of *current* was based on the the year at which the company was founded and a range of two year before and two year after the founded year was considered for calculation of *current* feature. A five year range including the current founded year is used for calculating the top sectors. Figure 4.11 demonstrates the example of how the ranges for past, current and future are calculated from the *founded year*. Based on these ranges the final table shows the count of top sectors for each company. For the calculation of the final table, from the given range we count all the companies that have invested in the top sectors and add it to the calculated feature *Top current sector*. Similarly, the calculation of *past* includes all the years before the selected current range and the count of all the sectors in which the company invested is calculated for the feature *Top past sector*. For the feature *Top future sector*, we include all the years after the selected current range to count the top future sectors. The past, current and future are the ranges given based on the year the company was founded. These segregation of sectors provides an insight about the shifts in the market and highlights the market trends of the company with each passing year. Knowing the trends in the past, current and future sectors not only provides an advantage to the investors but also to the entrepreneurs or decision makers to keep up with the trending market and invest wisely. Another calculated feature *Funding frequency* denotes the frequency of funding received by each sector

in the company. The calculation formula is given by :

$$\text{Funding Frequency} = \frac{\#\ of\ Sectors}{\#\ of\ Companies} \tag{4.1}$$

For example how many times the company received funding for software sector. These calculated features also helps to answer the question how many companies invested in the top sectors ? Investors and entrepreneurs benefit from this information as it helps them make an effective decision about whether to move, hold or sell their investment with respect to the changes in the market.

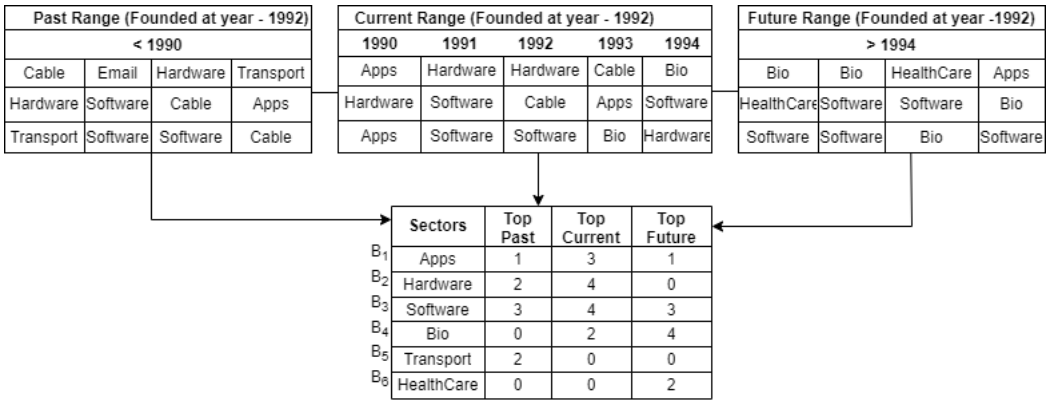


Figure 4.11: An example of creating past, current, and future ranges for market features. The top-middle table refers to top business sections at the "current" (*i.e.* 1990-1994). The top-left table refers to top business sections during the "past" (*i.e.* before 1990). The right-left table refers to top business sections in the "future" (*i.e.* after 1994). The table at the bottom shows number of times a sector tag appears in the companies with respect to "past", "current", and "future", respectively. For example, hardware tag appeared in two sponsored companies in the past, and appeared in four sponsored companies at the current.

4.1.4 Proposed framework

In this section, we briefly describe the structure of our proposed framework for business success prediction model as a binary classification task. For this, we use the two

Table 4.3: A description of Market Features used in the prediction model

Feature Names	Description	Type of Feature	Dimension Size after Encoding
Business sector	The market sector in which the business operates (for example Technology, Retail, Finance *etc.*)	Categorical	Tokenize using Count Vectorizer -480 dimensions
Top past sector	Calculated using count of top sectors in the past range of founded year	Numerical	1-D
Top current sector	Calculated using count of top sectors in the current range of founded year	Numerical	1-D
Top future sector	Calculated using count of top sectors in the future range of founded year	Numerical	1-D
Funding Frequency	Calculated using count of each sector divided by total number of companies	Numerical	1-D

primary database files : company and investor file. These two files contains complete information about companies and investors. In order to merge these files together to create our dataset for companies a unique identifier is used known a *Permalink* that joins the two files. In the next step, feature engineering and feature selection task is performed to extract meaningful features using our IBM triangle framework. Investor, business and market features are extracted, transformed and encoded for our final modeling task. The textual or categorical features such as Business Sector are extracted using Count Vectorizer technique which preserves the semantic meaning of the words used in the feature set. As Business sector column in the dataset includes multiple sectors in which the company operates in each row, Count Vectorizer method counts the number of times the word occurs within the dataset and assigns majority count to those words with maximum occurrences. The other categorical features such as the Funding Round Type, Country code etc, are either one hot encoded or label encoded depending on the size and dimension of the features. Based on the available set of feature new features are created to support and provide additional information for modeling and predicting business success. With this available feature set, we create a final dataset which includes the concatenation of investor, business and market features tied up to each company instances as shown in the Figure 4.12. The final dataset consists of a total of 526 features which is then used for modeling and predicting business success. In the next step, we apply different machine learning classification algorithms to the final dataset and compare the results using the

business target which is defined as :

$$y = \begin{cases} 1, \text{ If company status = “IPO” or “Acquired“} \\ \\ 0, \text{ If company status = “Closed”} \end{cases} \tag{4.2}$$

where y is the prediction value generated as the outcome of the modeling result. The business target determines the success and failure of the company.

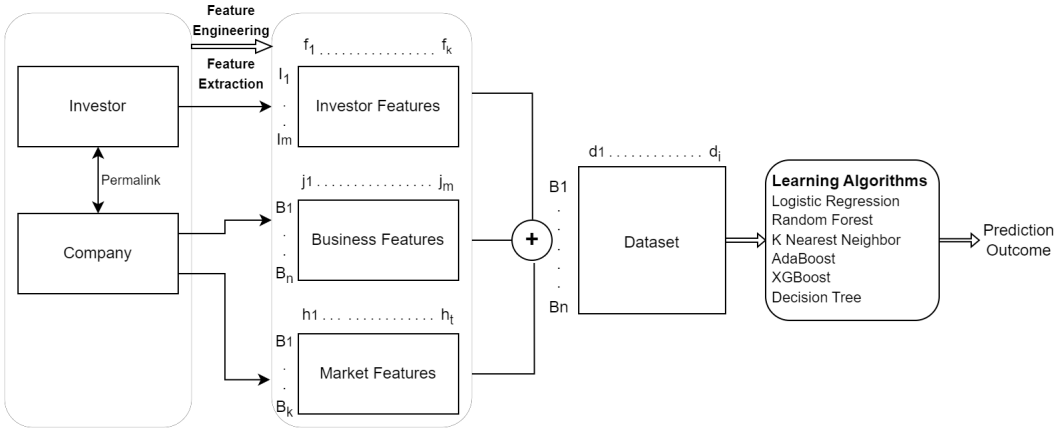


Figure 4.12: The proposed system flow chart for business success prediction. The original dataset has two tables: Company and Investor, linked by ”Permalink”. We first create investor features using Investor table. Business features and market features are derived from the Company table. The three features are consolidated to form IBM features to represent each business for learning and prediction

4.2 EXPERIMENTS

In this section, we first describe the benchmark data and the experimentation settings and then report the results and comparisons of several methods on the benchmark dataset.

4.2.1 Benchmark Dataset

We used two dataset in the experiments. The first dataset is Company file which includes information about companies, its demographics, sector, funding information

etc. The second dataset is Investment file which includes all aspects of Investments and investor details linked to the companies. These dataset is originally taken from a public data sources [1]. This dataset is extracted from Crunchbase.com containing 65k+ company details from 1800's to 2015. The original dataset consisted of four files "Companies", "Investments", "Acquisition" and "Rounds", out of which majority of the information regarding the business demographics, market sector and funding information were available in the company file and the investor information including the investor demographics, funding round code and funding round type were available in the investment file. Hence we chose company and investment file as the main data source and merged them together using a unique identifier *(Permalink)* to extract meaningful features from the two files given in the Table 4.4. In the next step, we use feature extraction step to include meaningful variables from the company and the investor files. We then perform a feature selection method on the selected variables using Chi-Squared and Pearson's correlation methods to extract top categorical and numerical variables from the benchmark dataset as shown in Figure 4.13. Out of the top features extracted, we selected top 12 features to create new dependent features. A total of 500 features we selected in the benchmark dataset to improve the model's performance.

After feature engineering, feature extraction, and removing all companies having the status of "operating" for accurate processing of business target, the final benchmark dataset is shown in Table 4.5. In this dataset the last column is our target variable which indicates whether the business will be successful (1) or not (0). The final dataset includes 13,334 records out of which training set consists of 10,668 records and testing set consists of 2,666 records divided into number of success and number of failed records with features representing each company information tied up to investor and market information as well.

[1] `https://github.com/chenchenpan/Predict-Success-of-Startups`

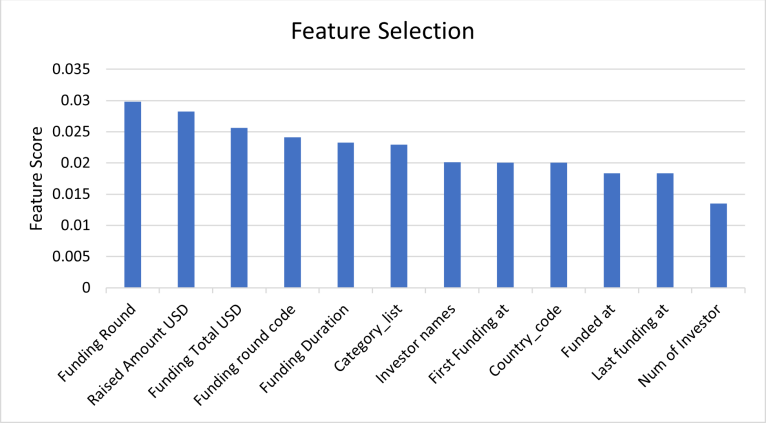


Figure 4.13: Feature selection method on benchmark dataset includes top 12 features based on the importance score to create new dependent features for modeling

Table 4.4: Simple statistics of the benchmark dataset. # "Companies" dataset lists all businesses. "Investments" dataset lists all investments investors made to the businesses

Data	# of Fields	# of Records
Companies	14	66,368
Investments	18	168,647

Table 4.5: Training *vs.* test split and respective class distributions (5-fold cross-validation was employed in the experiments. This table shows split of one fold)

Dataset	# of Successful	# of Failed	Total # of Instances
Train set	5,726	4,942	10,668
Test set	1,370	1,296	2,666

4.2.2 Comparative method

For our baseline method, we use the dataset obtained from company file and extract features related to company. No new features were added or modified from the investor data file as most of the previous studies used company data file for their

analysis [18, 99, 184, 188]. Hence for comparison purpose we keep our baseline similar to the previous studies and chose company features for predicting business success. The table 4.2.2 demonstrates the total features selected as baseline and their dimensions after encoding. A total of 500 features after conversion were used as a baseline method for predicting business success. In our proposed method for predicting business success, we selected features from company file and investor file and created a dataset which includes features concatenated from both the files. Based on this dataset, we created additional features related to our IBM framework to study and evaluate how different features influence the modeling results. A toal of 563 features were selected for modeling purposes in our proposed framework.

For fair comparison, all the experiments were performed on the same training and testing data with same number of instances. Eight machine learning classification algorithms were used with same business target of success or failure(0 or 1) for our baseline as well as for our proposed method.

- Logistic Regression is most commonly used model for binary classification tasks and has been used in many previous researches for predicting business success [18, 133]. However logistic regression has been known to have low performance as compared to tree-based algorithms.

- Random Forest has been known to achieve higher accuracy and is robust to noisy data as shown in previous studies [80]. For our experiments, we use Random forest with 100 trees and 200 trees for comparison of results.

- Decision Tree is a straightforward classification algorithm that produces comparable results for prediction tasks [132].

- K-Nearest Neighbor works by finding similar things in close proximity to each other. Hence in dataset, we use KNN as it helps to find similar companies for entrepreneurs or investors to compare and make decisions whether to invest or

not. KNN has not been explored much in the field of business prediction when compared to other machine learning models.

- XGBoost is a boosting technique that has gained tremendous popularity due to its high performance and enhanced speed in prediction tasks.
- AdaBoost [58] is another important boosting algorithms that have shown success in variety of machine learning applications such as bankruptcy prediction [202], failure prediction *etc.*
- Neural Network has been widely used for classification and regression problems due to its ability to offer better consistency and work in parallel to save processing time.

Table 4.6: Baseline Features to predict business success

Feature Names	**Type of Feature**	**Dimension Size after Encoding**
Business Sector	Categorical	Count Vectorizer 480- Dimensions
Company Status	Categorical	Binary label (0 or 1) 1-D
Country_code	Categorical	label encoded 1-D
Average Funding Duration Days	Numerical	1-D
Average Funding Duration Years	Numerical	1-D
Funding Rounds	Numerical	1-D
Funding Total $USD,$m,$b,$k	Numerical	3-D
Funding Duration Days, Month, Year	Numerical	3-D
Founded at Day, Month, Year	Numerical	3-D
First Funding Day, Month, Year	Numerical	3-D
Last Funding Day, Month, Year	Numerical	3-D

4.2.3 Experimental Settings

We implemented our experiments using the benchmark dataset provided in Table 4.5. A total of 13,334 instances for each company were selected after the Feature extraction, selection, and preprocessing stages. A total of 563 features related to Investor-Business-market features as shown in Table 4.1, 4.2, 4.3 were used for modeling and predicting business success for our proposed method and 500 features were selected for baseline method. Eight classification learning algorithms were applied in our experiments, including Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbor (KNN), Random forest with 100 trees (RF-100), Random Forest with 200 trees (RF-200), Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost) and Sequential Neural Network (NN).

All models were built using keras and scikit-learn library in python. For training the models, we separate the dataset into two portions, training and test set using 5-fold cross-validation, where (k-1) 4-folds are used for training and 1-fold is used for testing. In the preprocessing step, textual features such as *business sector* are converted into vectors and other categorical features are converted using either one-hot encoding or label encoding process depending on the size of the feature dimension. *Business Sector* is represented as textual features due to the fact that each company can operate on multiple industry sectors instead of one. Hence we use Count Vectorizer technique for conversion.

All results are obtained via 5 repeats of 5-fold cross-validation and our experiments are carried out on the training dataset and evaluated on the testing data.

4.2.4 Evaluation Metrics

To evaluate the quality of our prediction, we use average of area under the ROC curve as the main evaluation metric as it shows the accuracy of the binary classification model by ranking positive classes against the negative ones. In addition to this, we

also employ average accuracy as another performance indicator for assessing binary classification task. We use average accuracy for estimating classifier performance since we model and forecast business success using 5 repeats of 5-fold cross-validation. Using multiple 5-fold cross-validation separates the data into 5 equal-sized blocks and repeats the process 5 times. This helps in preventing the model from any kind of bias and over-fitting.

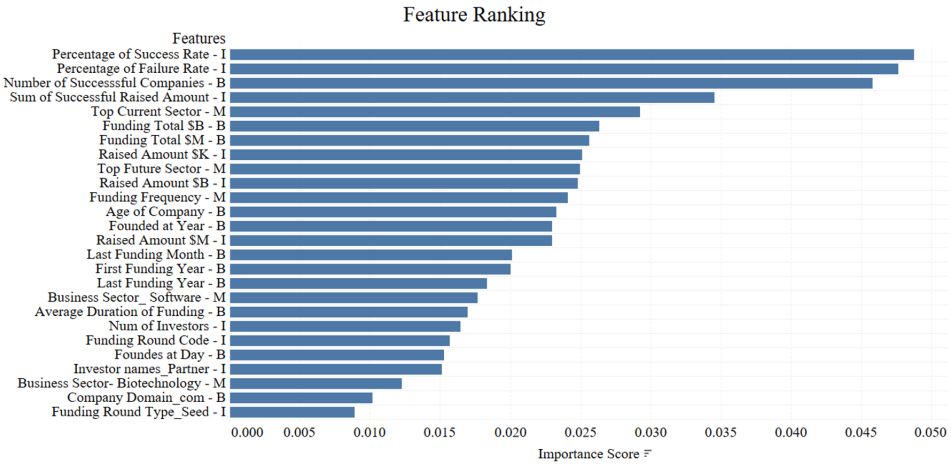


Figure 4.14: Ranking top features based on the importance score

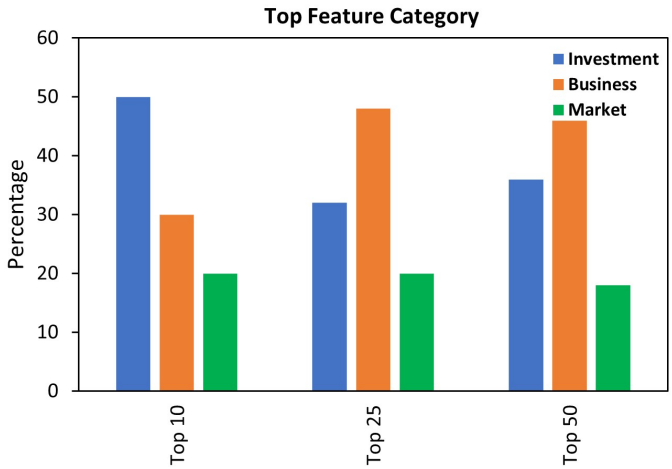


Figure 4.15: Percentage of top features in Investment, Business and Market angle

4.2.5 Experiment Results

Table 4.7 demonstrates the experiment results of our proposed model as compared to the baseline. Seven supervised learning algorithms were used along with neural network for business success prediction. For comparison purposes, we use the same experimental settings for baseline as well as our proposed method. Based on the results, it is observed that our proposed method of including IBM related features demonstrated improved performance when compared to the baseline. Hence, when predicting business success, it is important to consider features related to these business angles which is investment, business and market. To highlight the importance of selecting apropriate features ,we performed feature ranking method using the random forest classifier to train the model and get importance score for all the features. As shown in Fig 4.14, we flagged the feature category by I,B or M indicating the category in which the feature belongs. The top features based on the score includes *Percentage of success rate, Number of successful companies, Top Future sector, Total raised amount, Age of the company etc.* which highlights the importance of IBM in business success prediction. Apart from this, we demonstrate the shifts in the IBM features when selecting top 10, top 25 and top 50 features demonstrating the role of each feature category during the feature ranking process. As shown in Figure 4.15, Investment features cover 50% when selecting top 10 features whereas Business features remains on the top when selecting top 25 and top 50 features. Hence by considering our proposed method, the results demonstrates significant improvement when compared to the baseline.

Out of all the algorithms used for modeling, the Random Forest model and XGBoost model outperformed other classification algorithms used for modeling. Random Forest with 200 trees achieved the best average accuracy of 77% and mean AUC of 85%. The second best results were obtained from XGBoost model with average accuracy of 76% and AUC of 85% followed by other models. Although accuracy of

Table 4.7: Business success prediction results

Method	**Algorithm**	**Accuracy**	**AUC**
Baseline method - 500 dimensions	Logistic Regression	0.57	0.76
	Random Forest-100	0.74	0.81
	Random Forest-200	0.74	0.82
	Decision Tree	0.66	0.67
	K-Nearest Neighbor	0.69	0.75
	XGBoost	0.74	0.82
	AdaBoost	0.73	0.80
	Neural Network(sequential)	0.53	0.67
IBM Interplay Features - 563 dimensions	Logistic Regression	0.75	0.81
	Random Forest-100	0.77	0.85
	Random Forest-200	0.77	0.85
	Decision Tree	0.71	0.71
	K-Nearest Neighbor	0.69	0.75
	XGBoost	0.76	0.85
	AdaBoost	0.76	0.83
	Neural Network(sequential)	0.52	0.60

logistic regression and Neural network for baseline is low as compared to other models due to the fact that LR assumes linearity between dependant and independent variables [153]. Figure 4.17 shows the comparison of average accuracy for all models in baseline as well as our proposed method. Random Forest have been known to perform best in binary prediction task [116] due to it's robust nature and efficiency in handling small and large datasets. Boosting algorithms like XGBoost and AdaBoost have recently gained popularity to due it's execution speed and high performance.

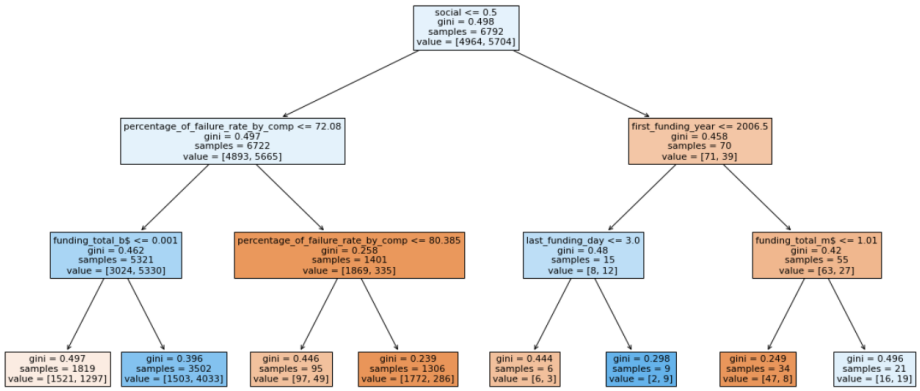
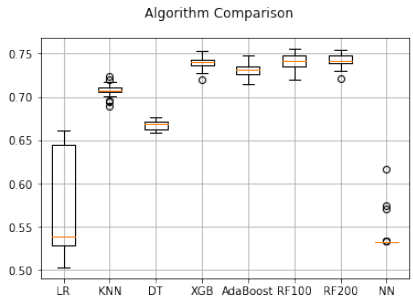


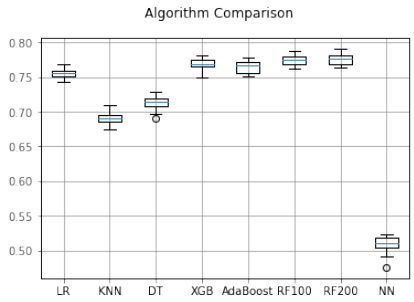
Figure 4.16: An example of a decision tree from Random Forest learned from proposed IBM features

Hence the results of Random forest are comparable to the boosting methods. Figure 4.16 demonstrates a snapshot of random forest tree with maximum depth of 3 and the first decision tree out of 200 trees estimator for simplicity of viewing.

In order to examine the difference between the baseline and using IBM features, Figure 4.18 reports the mean ROC curves and AUC values of all models in baseline as well as our proposed model. The ROC curve is useful as it helps to understand the trade-off between the True positive and False positive ratio. We can observe an improvement of 3% in Random Forest AUC. XGBoost model has shown an improvement of 4% and overall all the algorithms have shown some improvement except for KNN. Since KNN works by identifying similar patterns, it has not been widely used in business prediction or financial analysis, hence the results may vary from other algorithms. There is a slight 0.001% difference in the accuracy of KNN algorithm baseline and proposed method. The best performance in terms of AUC is obtained by XGBoost which is followed by Random forest.

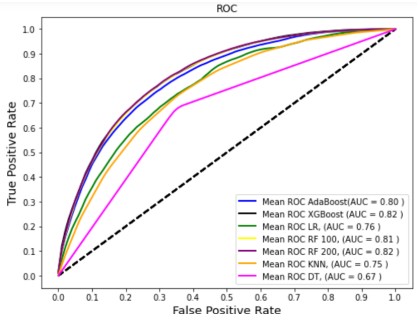


(a) Average accuracy of baseline method

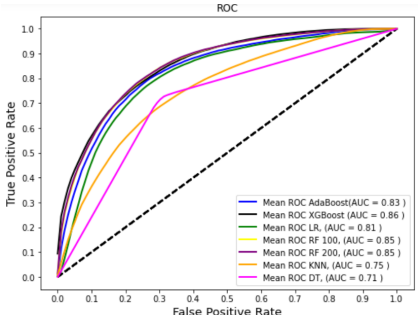


(b) Average accuracy of the proposed method

Figure 4.17: Comparisons of average accuracy of all seven learning algorithms



(a) Average AUC for baseline method



(b) Average AUC for proposed method

Figure 4.18: Comparison of ROC curve and AUC values for all models

4.3 CONCLUSION

In this research, we proposed an triangular framework known as IBM triangle for business success prediction and highlighted the importance of using Investor-Business-Market interrelations to identify critical features that makes a huge impact in predicting business success. Based on the study, we highlighted the importance of each of the IBM entity and including these features for business prediction significantly improves the performance of the model. Following the proposed triangular feature

relationship, we elaborated on the technical details of extracting these features from the benchmark dataset. Based on the available set of features and keeping the IBM framework in place, we created additional features to enhance the performance of the model. Seven supervised learning algorithms are applied to the datasets by using new IBM features. Experiments and comparisons confirm that IBM feature-based methods not only outperform baseline methods in predicting business success but also provide a meaningful and transparent understanding of feature importance in the prediction. This research validates the effectiveness of computational methods, combined with carefully designed features, in the modeling and prediction of business success.

CHAPTER 5
GRAPH LEARNING MODELS FOR BUSINESS SUCCESS PREDICTION

Over the past years, there has been a signification amount of research in predicting business success using quantitative and qualitative features [35, 49, 66, 164] with a focus on supervised machine learning model used for prediction task. There have been significant improvements in results, and researchers have progressed in their studies by analyzing different business angles for success prediction [61]. However, very few studies have been made by focusing on the topological features of business success [23, 65, 200]. The topological features include deep relations between different entities that identify interaction between different objects and further construct a network for predicting business success. This type of structure provides a different perspective to the audience when predicting business success. Hence, to further one step ahead in our study, we propose a heterogeneous graph attention network that demonstrates the relationship between different types of entities. Our ultimate goal is to predict the success of the business by analyzing different features and relations between business-related entities.

Many real-world applications use Heterogeneous graph learning, which is ubiquitous to real-world scenarios. For example e-commerce platforms usually have (user-item-vendor) relations [198], similarly healthcare sector have 3 types of relation (patient-disease-drug) [172]. Another example that demonstrates three types of relation is a bibliographic network representing (author-write-paper) relation using four types of entities.

In a heterogeneous graph network, entities demonstrate a unique type of rela-

tions which is not commonly observed unless we dig deep into it. For example in e-commerce sector relationship between user-product or product-vendor are commonly identified as they a bipartite relations, however, observing relationship such as user-product-vendor is often missed as a particular user may be specific to using a product from a particular brand [198]. Hence, it is important to have an accurate link prediction setting to understand and capture such intricate relations. Another important thing is the attention mechanism in heterogeneous graph which provides more critical attention to the neighboring nodes and edges and assigns a score for each entity such that no important information is missed. Although heterogeneous graph has it own advantages, it is challenging to capture such type of relations and extract meaningful nodes and edges out of the network.

For business success prediction, we identify four nodes with three types of relation between them such as (person-company-investor) Identifying such relations are complex which also creates a challenge for existing embedding methods to be directly applied to the heterogeneous graph network. The below paragraph we highlight the challenges of heterogeneous graph structure:

- Imbalanced Nodes: Heterogeneous graphs have multiple node types, and often have severely imbalanced nodes where one type of node is often much more than other types (e.g. number of Investors are far more than the number of companies). As a result, traditional graph methods cannot be directly applied as most of the graph algorithms are based on homogeneous properties.

- Imbalanced network edges: In many heterogeneous networks, edges between two types of nodes, such as Investors connected with person or person connected with a market sector are rare events, which makes it difficult to learn effective features or consider features between these types of edges.

Motivated by the above observations, in this study, we propose a HAN (Hierarchical attention network) model which takes the company, investor and person dataset in

a tabular format and converts into a heterogeneous graph format with four nodes and edge connections between them. Feature engineering and feature learning method gathers relevant features for each node and also captures implicit relationships between them. We define meta-paths which is a sequence of node type and edge type that describes intricate patterns between the nodes. These meta paths and selected features are provided as input to the model enabling more nuanced insights and a more sophisticated network analysis approach. Initial experimental results demonstrated that the HAN model outperforms the traditional machine learning model for predicting business success.

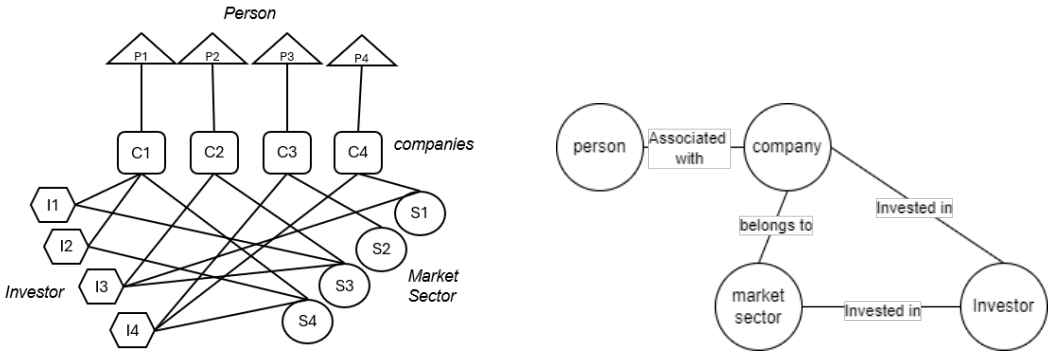
5.1 PROBLEM DEFINITION

Heterogeneous Graphs: In this chapter, we define heterogeneous graphs as : $G = (V, E, R, T)$, where $V = \{v_i\}_{i=1,\cdots,N}$ is a vertex set representing the nodes in a graph, and $e_{i,j} = (v_i, r, v_j) \in E$ is set of edges connecting different nodes and each edge may have different relationship type. T is a set of node types $T(v_i)$. In our business prediction problem the node set V consists of four node types $V = V_1, V_2, V_3, V_4$ which in turn encompasses into four types of entities in a heterogeneous network, such as Company node (V_1), Person node (V_2), Sector node (V_3) and Investor node (V_4).

In the heterogeneous network, dealing with different types of nodes brings in a semantic relationship between nodes and edges. This complexity is generally described using meta-data modeling through the network schema and meta-paths.

Formally, a meta-path is a sequence of nodes and edges that shows different types of relationships between node types as: $A_0 \xrightarrow{R_0} A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_k} A_{k+1}$ describes a composite relation between node type A_0 and A_{k+1}. For example, two companies can be connected via multiple meta paths (Company-Person-Company) or (company-investor-company). Hence, the meta path shows complex relations between entities. Figure 5.1 shows an example of the heterogeneous graph representing a business net-

work. Different meta-paths represent different semantic information, such as the company belongs to a sector which is invested by an investor "(CSI)". This information is useful for capturing rich semantics between different nodes.



(a) Heterogeneous example of company network using four types of nodes and interconnected edges

(b) Heterogeneous network schema

Figure 5.1: Heterogeneous Network representation of business schema

5.2 THE PROPOSED METHOD

In this section, we first describe the HAN network and then propose a hierarchical graph attention network (HAN) framework to predict business success using a heterogeneous graph structure

5.2.0.1 HAN model

A heterogeneous graph is a special kind of graph that contains either multiple types of nodes or multiple types of edges. The complexity of handling such structural information, as well as preserving its semantic meaning, is well handled by HAN architecture. One of the key aspects of the HAN model is its attention level mechanism, which is proposed to learn the importance of nodes and their neighbors to develop a node classification problem.

Node-level Attention Node-level attention in the HAN network aims to learn information from meta-path-based neighbors for each node and then aggregate such information to represent the meaningful neighbors for a node embedding representation. Given a node v_i and the neighbor of the node $N(v_i)$, the attention mechanism is applied to select the most important neighbor and assign a score to it. The mathematical representation of node-level attention is given in two parts:

- First, we project the features of nodes and their neighbors into a shared space using a transformation matrix denoted as:

$$\mathbf{W}\epsilon\mathbb{R}^{d\prime *}d \tag{5.1}$$

  where $z_i = Wh_i$ and $z_j = Wh_j$. Here z_i is the projected feature vector, and $d^{'}$ is the dimension of the projected space.

- Next, we calculate the attention coefficient a_{ij} that represents how important the node v_j is to v_i using shared attention mechanism.

$$e_{ij} = LeakyReLU(a^T[z_i||z_j]) \tag{5.2}$$

  where $a\epsilon\mathbb{R}^2d'$ is a weight vector, and LeakyReLU is the activation function.

- In order to normalize the attention score e_{ij}, we use the equation below, which normalizes the score across all neighbors using the Softmax function

$$a_{ij} = \frac{exp(e_{ij})}{\Sigma_K\epsilon_{N(v_i)}exp(e_{ik})} \tag{5.3}$$

- The final node representation $h_i^{'}$ of the projected neighbor node is obtained by aggregating features for node embedding, weighted by the attention score a_{ij} :

$$h_i^{'} = \sigma\left(\sum_{j\epsilon N(v_i)} a_{ij}h_j\right) \tag{5.4}$$

where σ denotes the activation function like ReLU. Once we analyze the node-level attention, we identify meta-paths and aggregate the meta-path-based attention. The mechanism of calculation is similar to node-level attention; hence, we skip the mathematical representation. In the final model input, the combination of node-level attention and meta-path-based aggregation is applied to the HAN model to learn meaningful representations of different entities in a heterogeneous graph format. Table 5.1 shows the mathematical representation of symbols with their description.

Table 5.1: Mathematical representation of notations used in HAN model

Symbol	Description
$G = V, E$	Graph with nodes V and edges E
T	Set of node types
R	R is relationship of object G where $R = R_0, R_1, \ldots R_k$
A	$A = A_0, A_1, \ldots A_k$
P	Set of meta Paths $P = A_1, A_2 \ldots A_k$
a_{ij}	Attention coefficient
σ	Activation function
h_i	Feature vector of node v_i
d'	Dimension of node
z_i	Transformed or projected feature vector
W	Transformation Matrix

Meta-path construction for Business-related Nodes Meta-paths are a sequence of relations that describe how two nodes are connected to each other. A Meta-path generates possible connections between the nodes by defining a path

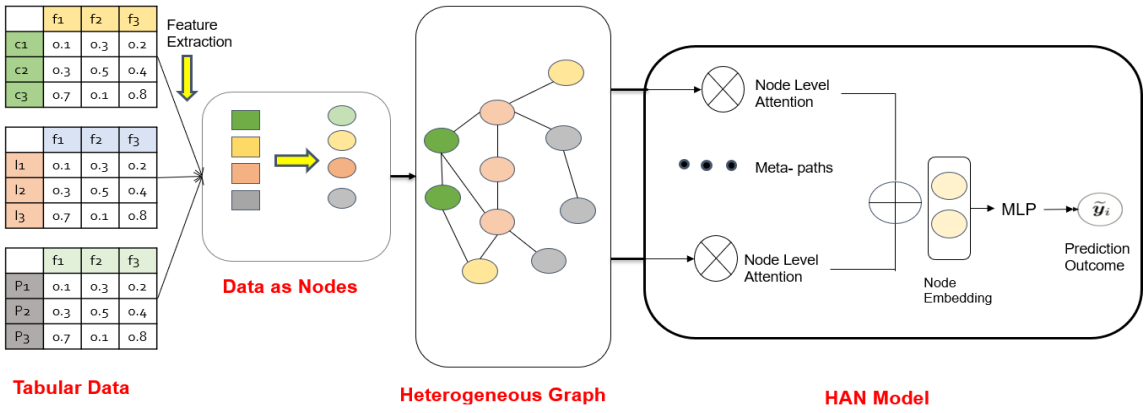


Figure 5.2: The overall architecture of the proposed Hierarchical Attention Network for node classification task to predict business success

that connects them together. For our Company data, we identify four types of nodes: Company(C), Investor(I), Person(P) and Sector(S). Each of these is connected via an edge between them. For our proposed model, we include only direct connections to generate meta-paths between the nodes. We identified 6 meta paths for business modeling : Investor invests in Company and belongs to a particular sector (I-C-S), Company is associated with a person and belongs to a sector (C-P-S), Company is associated with an Investor who invests in a Sector (C-I-S), A person works for a company and is associated with a sector (P-C-S), A company employs a person who is also an investor (C-P-I), A company is associated with an Investor who has invested in a company (C-I-C). These relations provide a deeper meaning between the nodes and their interconnection to each other. Such type of information is important when predicting business success as it provides a new way of looking into the features with a new business angle to generate additional features for binary classification tasks.

5.2.1 Heterogeneous Graph Attention Network Framework

The Figure 5.2 shows the framework architecture of the HAN model. The first step is the feature extraction process, for this, we extract the information from three data files Business, Investor and Person depicting information about the companies, investors details and person information available in a tabular format. These three datasets are extracted to capture meaningful relations about the company, investor, and person associated with the company. The main objective is to utilize the available relations and create additional feature dimensions to improve the prediction results. In the feature extraction process, we include all the features described in the previous chapter, including additional features created using the IBM framework for our prediction framework, and add a person dimension to the extracted company data. Once the dataset is ready for processing, we then build a graph model for prediction. In order to do so, we convert the tabular data into a heterogeneous graph format with nodes and edges representing the relation between them. We extracted four types of nodes from the Investor, company, and person files, which in turn provided us with the company node, investor node, person node, and sector node. These four nodes represent a strong relationship with each other and are interconnected via one or more edges. For example, a person node is connected to one or many companies, and an investor can be connected to one or more sectors. However, there is also a case where the person node has no connection with the investor node or sector node. Hence, with heterogeneous properties, we can capture such complex semantics to determine strong and weak connections between the nodes. In the next step, we identify meta paths from the heterogeneous graph structure to capture rich semantics between the nodes. The meta paths highlight how two nodes can have different relationship types, demonstrating the importance of capturing each relationship type to extract useful features for the link prediction task. The main step of our proposed framework is the attention mechanism in heterogeneous graphs. Before capturing any information

from meta-path neighbors for each node, it is important to note that each node plays a different role to its neighbor and hence shows different importance in the learning node embedding task. For example, the person node has no direct link with the sector node but is strongly connected to the company node. For this, node-level attention plays an important role in capturing the importance of meta-path-based neighbors for each node. We capture the node-level attention using different meta-paths generated and analyze the most important meta-path-based neighbors, and assign a score to it. For each node, node-level attention aim to learn the importance of meta-path based neighbor and assign different score to them. The highest score is given to the most important neighbor, which identifies a close relation between the nodes and the features. This process enhances feature selection and offers new business angle apart from our IBM triangle for further exploration of business success. Apart from this, heterogeneous property provides us with importance of different nodes at varied stages which is highly important for businesses at different stages of their cycle.

In the final step, we embed the available data into the HAN model to predict business success, including nodes, edges, meta-paths, features, and the target variable.

5.3 EXPERIMENT

5.3.1 Benchmark Dataset and Heterogeneous Data construction

For our benchmark dataset, we extracted data from Crunchbase.com using Crunchbase API. These files include Company profiles, Person profiles, Investor and funding profiles. From these data sources, we used company, person, and investor files as they included the majority of the information about business-related features needed for the prediction task. Figure 5.3 shows the dataset of company, investor and person files with features extracted for modeling. Overall, the dataset consists of 101,049 companies, 133,394 persons, and 168,647 investors. From this, we removed companies that were in "operating" status. In the final dataset, we create company, investor,

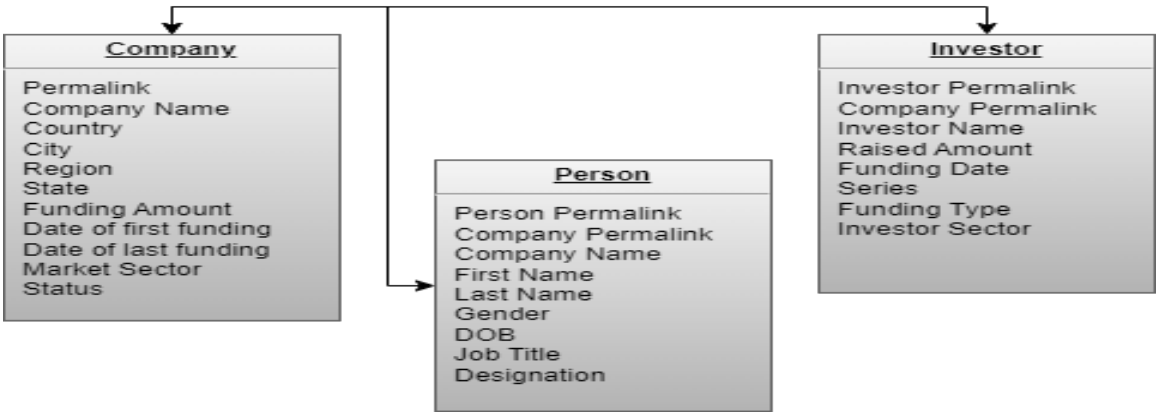


Figure 5.3: Tabular data for Company, Investor and Person taken from Crunchbase platform

and person datasets, each linked to one another with a unique identifier known as "Permalink".

In order to convert the tabular data into heterogeneous graph format, we extract four types of nodes (Company, Investor, Person and Sector) from three tabular files mentioned above and edges (Company-Investor), (Company-Person), (Company-Sector), (Investor-Person) and (Investor-Sector) from different node entities that form meaningful relations with each other. We also extract features from Company, Investor, and Person files, which include domain information, funding information, Person details such as Person name, role, designation, etc., and Sector information such as IT, Health, Manufacturing, and label for each company. The Table 5.2 shows the heterogeneous data after conversion. We also add features to the previous study's extracted information to elaborate our research and provide advancements in our proposed problem.

This heterogeneous graph data along with features and label then becomes our input in the final model for prediction.

Table 5.2: Dataset for Heterogeneous graph with number of nodes and edges for each entity

Dataset	# of Nodes	Relations	# of Edges	# Features	# of meta paths
Crunchbase: Company, Investor, Person Data	Company - 790	Company-Investor	Total number of edges = 24,506	245 – After one hot encoding	6
	Investor - 613	Company-People			
	Person - 1086	Company-Category			
	Category - 151	Investor- Category			

5.3.2 Experimental settings

For the Proposed HAN model, we use the input generated in a heterogeneous format with node embedding, meta-paths aggregation, and node-level attention mechanism to learn the importance of each node and its neighbors. Finally, a fully connected layer is applied to predict the output of business success as a binary classification task with 1 or 0. For a binary classification task, we use a loss function as cross entropy and apply Adam optimizer to improve training speed and convergence. We use L2 as a regularization parameter to penalize large weights in the model. We set the learning rate as 0.01 to see how quickly a model can converge. Our model results with experimental setting are shown in Table 5.3. We use an early stopping patient of 100, where the model stops training if the validation error does not decrease. Our dataset is split into training and test sets with 80% for training and 20% for testing using random split. We also apply different hyperparameter tuning by changing the number of attention heads, learning rate, and number of meta-paths as input to adjust the best settings for our model. The best results were obtained using the setting as described above. All our experiments are performed using Pytorch geometric framework for graph modeling due to its flexibility and adaptability for deep neural network models.

In our baseline comparison, we chose the Random Forest model with 200 trees as we got the best results in our previous study with this model. We keep the same

Table 5.3: Results of HAN model compared to Baseline method

	Model	Setting	Test Accuracy	# of Classes	Meta paths
Baseline Method	Random Forest- 200 trees	Train-test split 80 - 20	74%		
HAN Model	Input – Meta paths, Num of features, Hetero data with nodes and edges, num of classes	Train-test split 80 - 20	83%	2	Company-Investor-Company
					Company-Category-Company
					Company-Person-Company
					Investor-Category-Investor
					Investor-Company-Investor
					Person-Company-Person

experimentation settings, such as the train-test split ratio and number of features generated, to ensure fairness and comparative results.

5.3.3 Experimental Results

For our experimentation purpose, we use HAN model architecture as our framework for our proposed method using Graph Neural Network (GNN) to predict business success. The final prediction output is a classification task with an outcome of 1 or 0. The results of our proposed model demonstrated improved performance, as shown in Table 5.3 when compared to the baseline method. HAN model has been used in many applications such as text classification, healthcare analytics, and social network structure and has shown tremendous improvements in the overall results. Hence, for our prediction task, applying HAN architecture was helpful as it takes node relations into consideration and extracts meaningful relations between the nodes. As our dataset is sparse and contains different nodes with different types of relations, it is important to do some statistical analysis to show how sparse the data is in terms of graphs. Figure 5.6 demonstrates the degree distribution of the entire network, including different nodes. We also highlight the degree distribution of the company and the investor's nodes in Figure 5.4, 5.5 with respect to the count of each node in the figure. The degree distribution is an important metric in network graphs which helps us understand the topology and the dynamics of the network. It gives us an

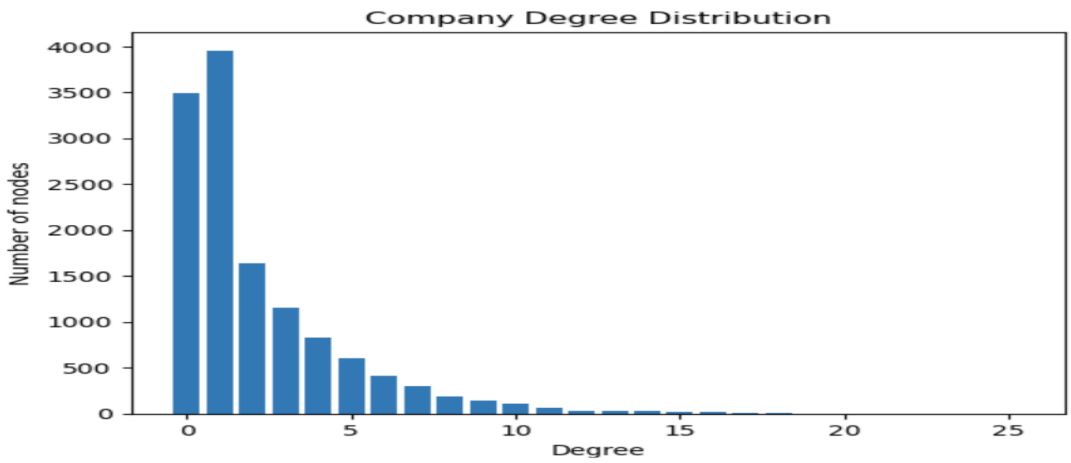


Figure 5.4: Degree distribution plot for Company Nodes

insight into how nodes are connected to each other and identify relevant patterns between the nodes. For our baseline, we chose the supervised learning method for the prediction task as it has been commonly applied for binary classification problems in various studies. We chose the Random Forest model for comparison due to its robust nature and flexibility in handling complex data structures. In our previous study (Chapter 4), we used several supervised learning model for business success prediction and compared their results, out of those models, Random Forest model with 200 trees performed the best in terms of accuracy and AUC score. Hence, for comparison purposes, we chose random forest as our baseline since it performed the best in our previous study and compared it with our proposed method. The results demonstrated superior performance with an improvement of 9%, showing that the HAN model has good interpretability in graph analysis.

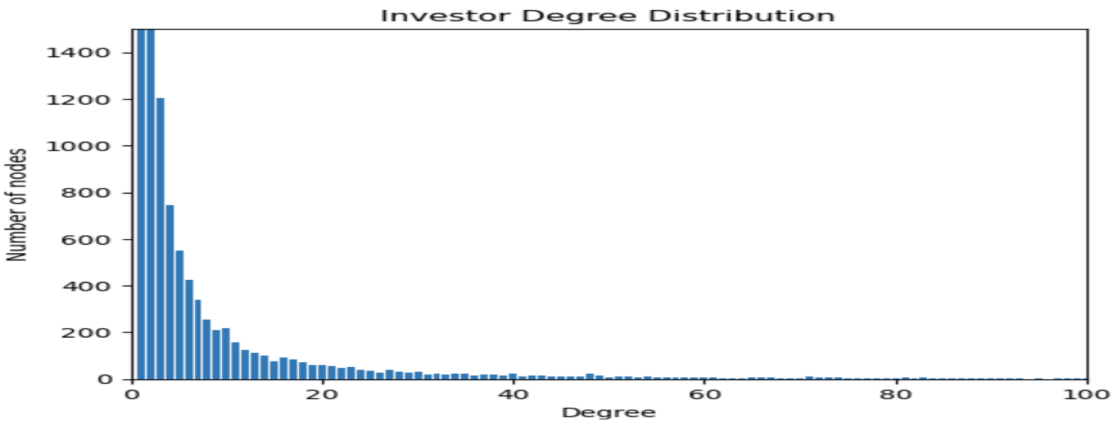


Figure 5.5: Degree distribution plot for Investor Nodes

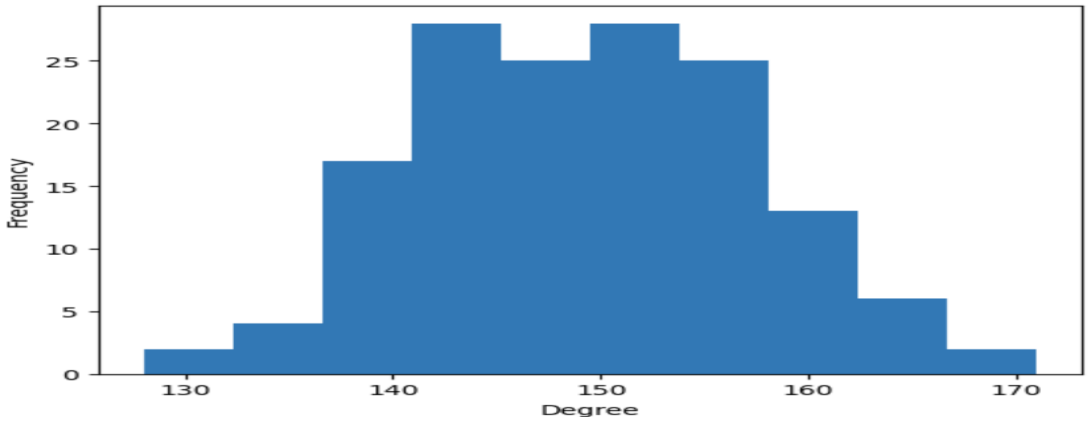


Figure 5.6: Degree distribution plot for the entire Network

5.4 CONCLUSION

In this chapter, we studied the heterogeneous graph attention network for business success prediction with a focus on the attention mechanism to capture the node-level attention of the neighbors in order to depict the importance of neighbors in our prediction task. The framework analyzes useful neighbors that are important for the prediction task and gives them the highest score based on the importance level. The HAN model is capable of capturing rich semantics between nodes and edges, leveraging node-level attention to learn the importance of different nodes and their meta-paths affecting different nodes. In this study, We used the HAN model on a real-world dataset to validate its performance on a business prediction task. Additionally, our study utilizes the knowledge from previous theories to extend the prediction to the next level. For this, we analyze the business from a new angle by including features of business, investor, person, and market sector and demonstrate how these relations affect the overall growth of a company. The HAN model utilizes the structured information and features represented in a uniform way to analyze the outcome of business. Finally, the experimentation results demonstrate that the HAN model is superior in the classification task of predicting business success when compared to the baseline method using the Random Forest model for the prediction task. By analyzing the knowledge from leveraging attention level mechanisms and meta-paths for business prediction, the HAN model has proven to be effective and demonstrated good interpretability in predicting business success. In the future work, we seek to investigate further as to how we can apply network models in other business applications.

CHAPTER 6

CONCLUSION & FUTURE DIRECTIONS

Machine Learning models are being widely used in many applications due to their capability to provide data-driven insights and the ability to identify useful patterns and trends. However, due to the increasing amount of data and potentially ever-changing time, the complexity of analyzing the data accurately also reduces as the data becomes obsolete with the changing times. Hence, in order to keep up with the changing times, there is a need for a combined system that can handle the dynamic environment for data management as well as create a robust model for prediction tasks.

In this thesis, in order to promote a robust predictive analytics method for business prediction tasks, we propose a framework that first analyzes different aspects of business based on the relevant theories and literature and then builds a machine-learning model for our prediction task. Our ultimate goal is to predict success of the business, for this we first need to come up with factors that are relevant for business success, then we analyze those factor to extract meaningful features for the business prediction task and finally we build a learning model for success prediction and evaluate the modeling results. The contributions and evaluation results of the proposed model are summarized as below :

The contributions and evaluation results of the proposed models are summarized as below.

6.1 CONCLUSION

1. Factors relevant to business success

- **Contributions.** 1) We first study the business theories and market to analyze the reason behind the rise and fall of the businesses. 2)we then distinguish the external and internal factors to make valid predictions on business success. Extracting these factors is a challenging task as businesses are constantly evolving due to competitive markets and the Industrial Revolution. 3)Hence, in order to overcome this challenge, we create a systematic framework based on relevant theories and studies that characterizes several factors into three main entities, namely, Investment, Business, and Market (IBM). Many established firms and startups have unique information available about their company, such as their financial information, investments, funding amount, innovations, employee details, etc. 4) We characterize this available information into three main entities for the prediction of business success using machine learning algorithms. These three key elements are primarily responsible for business success and are, therefore, interrelated to each other.

2. Quantifiable Features Used for Learning

- **Contributions.** 1)We first define the success of the business by measuring key factors that are responsible for business growth. These two key factors, that is, the capability of a company to become an IPO or have an M&A with a company of the same level or higher, become our indicator for success. 2) Based on these key factors and IBM entities, we then identify features responsible for predicting business. 3)Next, we perform feature engineering and a feature selection process to select the most appropriate features for learning and create additional features that are more informative and quantifiable for our prediction task. For example, from the original set of features provided in the dataset, such

as the "Number of investments" made by the investor within the company, we calculate the "percentage of success rate" and "percentage of failure rate." Similarly, from the "first funding date" and "last funding date", we calculate the" funding duration" of each company. We do this for every business angle, such as for investment features, market features, and business features to cover all aspects of business. These features are measurable and contribute to the growth of the business. 4) Finally, we provide a well fitted bias-free model to predict business success using these features as input. For our prediction task we include all types of firms such large- scale companies, startups, mid-size firms to help them make informed decisions and stay on top of the market.

3. Learning Model and Performance

- **Contributions.** 1) We propose two methods of learning to predict business success, including supervised machine learning algorithms for binary classification tasks based on the label generated and Graph Heterogeneous graph attention network(HAN) to identify semantic relationships between the nodes. The goal is to produce results that will be valuable for investors, entrepreneurs, stakeholders, and other researchers who are interested in taking this research one step ahead. 2) For supervised learning methods, We leverage the use of the triangular framework between investor, business and market and their interrelationship with each other to develop additional features as well as use the extracted features from the dataset for modeling business success. we use seven supervised learning algorithms for our learning task. The experiments performed on the business dataset demonstrate that considering features based on the triangular framework can indeed help in improving the accuracy of the model when compared to the baseline method, which is applied only to the available features of the company. 3) For a network-based model, we propose a

graph learning method to extract semantics relations between different entities through a heterogeneous graph model. The model is optimized to make accurate predictions based on its property to extract meta-paths and attention level mechanism which provides a new angle to business prediction. We use node-level attention to capture useful features and information about the important neighbors. The results demonstrate improved performance of the HAN model when compared to the baseline, which is the Random Forest Model.

6.2 FUTURE DIRECTIONS

For our business prediction problem, much of the focus was given in developing a framework based on three different business angles and exploring various factors in analyzing vast features that are measurable for predicting business success. Another limitation was the availability of the dataset from public platforms which released only limited information. As the dynamics of the business keeps changing rapidly, there is always a risk of accurately predicting whether the company was successful or failure as the prediction is based on the time of the data extraction process.

Future studies should explore different business angles based on the dataset extracted from various sources and combine them to extract more knowledge on business-related information. In future studies, also hope to get real-time data exploration and analysis to improve the existing methods for prediction. Moreover, nowadays there are deeper models available for prediction task with more flexibility and advancements, therefore it is important to study how we can utilize the available models for business prediction. With the recent advancements in AI and LLM models, there is a vast discussion on how AI can revolutionize business innovations and generate new business models for companies to explore and adapt to generate faster growth and success [29]. AI has improved several aspects of the business, including operations, RD, Sales, and customer interactions, and even generated new business features such as

ChatBot assist, personalized product recommendation, self-service portals, etc, which were not present earlier [84]. Several new business sectors have been established, and a few businesses that were booming in the past have now become inoperable. These features and advancements in Industry 4.0 have opened new doors for researchers to explore how AI can benefit in the prediction of business success. Hence, future research must also include AI models to analyze different aspects of business.

BIBLIOGRAPHY

[1] M. Abdullah. The implication of machine learning for financial solvency prediction: an empirical analysis on public listed companies of bangladesh. *J. of Asian Business and Economic Studies*, 2021.

[2] E. Akyildirim, A. Goncu, and A. Sensoy. Prediction of cryptocurrency returns using machine learning. *Annals of Operations Research*, 297:3–36, 2021.

[3] J. Al-Henzab, A. Tarhini, B. Y. Obeidat, et al. The associations among market orientation, technology orientation, entrepreneurial orientation and organizational performance. *Benchmarking: An Intl. J.*, 2018.

[4] A. Alamsyah and T. B. A. Nugroho. Predictive modelling for startup and investor relationship based on crowdfunding platform data. In *Journal of Physics: Conference Series*, volume 971, page 012002. IOP Publishing, 2018.

[5] R. Allu and V. N. R. Padmanabhuni. Predicting the success rate of a start-up using lstm with a swish activation function. *Journal of Control and Decision*, 9(3):355–363, 2022.

[6] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[7] M. Aminova and E. Marchi. The role of innovation on start-up failure vs. its success. *International Journal of Business Ethics and Governance*, pages 41–72, 2021.

[8] Y. Q. Ang, A. Chia, and S. Saghafian. *Using machine learning to demystify startups' funding, post-money valuation, and success*. Springer, 2022.

[9] Y. Q. Ang, A. Chia, and S. Saghafian. Using machine learning to demystify startups' funding, post-money valuation, and success. pages 271–296, 2022.

[10] T. Antretter, I. Blohm, D. Grichnik, and J. Wincent. Predicting new venture survival: A twitter-based machine learning approach to measuring online legitimacy. *J. of Business Venturing Insights*, 11:e00109, 2019.

[11] J. Arroyo, F. Corea, G. Jimenez-Diaz, and J. A. Recio-Garcia. Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, 7:124233–124243, 2019.

[12] W. B. Ashton and R. K. Sen. Using patent information in technology business planning—i. *Research-Technology Mgmt*, 31(6):42–46, 1988.

[13] M. M. Aung, T. T. Han, and S. M. Ko. Customer churn prediction using association rule mining. *International J. of Trend in Scientific R&D*, 3(5):1886–1890, 2019.

[14] M. Bangdiwala, Y. Mehta, S. Agrawal, and S. Ghane. Predicting success rate of startups using machine learning algorithms. In *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE, 2022.

[15] R. J. Barro. The stock market and investment. *The review of financial studies*, 3(1):115–131, 1990.

[16] J. J. Beckwith. Predicting success in equity crowdfunding. 2016.

[17] F. R. d. S. R. Bento. Predicting start-up success with machine learning. Master's thesis, Universidade NOVA de Lisboa (Portugal), 2017.

[18] F. R. d. S. R. Bento. *Predicting start-up success with machine learning*. PhD thesis, Universidade Nova, 2018.

[19] S. Bernstein, X. Giroud, and R. R. Townsend. The impact of venture capital monitoring. *The Journal of Finance*, 71(4):1591–1622, 2016.

[20] B. Bini and T. Mathew. Clustering & regression techniques for stock prediction. *Procedia Technology*, 24:1248–1255, 2016.

[21] V. Boasson and A. MacPherson. The role of geographic location in the financial and innovation performance of publicly traded pharmaceutical companies: empirical evidence from the untied states. *Environment and Planning A*, 33(8):1431–1444, 2001.

[22] M. Böhm, J. Weking, F. Fortunat, S. Müller, I. Welpe, and H. Krcmar. The business model dna: Towards an approach for predicting business model success. *Wirtschafts informatik*, 2017.

[23] M. Bonaventura, V. Ciotti, P. Panzarasa, S. Liverani, L. Lacasa, and V. Latora. Predicting success in the worldwide start-up network. *Scientific reports*, 10(1):345, 2020.

[24] J. Bonello, X. BrÉdart, and V. Vella. Machine learning models for predicting financial distress. *Journal of Research in Economics*, 2(2):174–185, 2018.

[25] P. Borchert, K. Coussement, A. De Caigny, and J. De Weerdt. Extending business failure prediction models with textual website content using deep learning. *European Journal. of Operational Research*, 2022.

[26] I. Bose and X. Chen. Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce*, 19(2):133–151, 2009.

[27] E. Bouri, C. K. M. Lau, B. Lucey, and D. Roubaud. Trading volume and the predictability of return and volatility in the cryptocurrency market. *Finance Research Letters*, 29:340–346, 2019.

[28] A. Breitzman, P. Thomas, and M. Cheney. Technological powerhouse or diluted competence: techniques for assessing mergers via patent analysis. *R&D Management*, 32(1):1–10, 2002.

[29] A. Brem, F. Giones, and M. Werle. The ai digital revolution in innovation: A conceptual framework of artificial intelligence technologies for the management of innovation. *IEEE Transactions on Engineering Management*, 70(2):770–776, 2021.

[30] A. B. Brik, B. Rettab, and K. Mellahi. Market orientation, corporate social responsibility & business performance. *J. of Business Ethics*, pages 307–324, 2011.

[31] J. Brüderl, P. Preisendörfer, and R. Ziegler. Survival chances of newly founded business organizations. *American sociological review*, pages 227–242, 1992.

[32] It's hard to say goodbye. a compilation of startup failure post-mortems by founders and investors, 2024.

[33] V. Cerra, A. Fatas, and S. C. Saxena. Hysteresis and business cycles. *IMF Working Papers*, 2020(073), 2020.

[34] G. N. Chandler and S. H. Hanks. Market attractiveness, resource-based capabilities, venture strategies, and venture performance. *Journal of business venturing*, 9(4):331–349, 1994.

[35] S. K. Chawla, C. Pullig, and F. D. Alexander. Critical success factors from an organizational life cycle perspective: Perceptions of small business owners from different business environments. *J. of Business and Entrepreneurship*, 9(1):47, 1997.

[36] M.-C. Chen, R.-J. Wang, and A.-P. Chen. An empirical study for the detection of corporate financial anomaly using outlier mining techniques. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, pages 612–617. IEEE, 2007.

[37] S. Cheriyan, S. Ibrahim, S. Mohanan, and S. Treesa. Intelligent sales prediction using machine learning techniques. In *Intl. Conf. on Computing, Electronics & Communications Engineering (iCCECE)*. IEEE, 2018.

[38] J. T. Chin. Location choice of new business establishments: Understanding the local context and neighborhood conditions in the united states. *Sustainability*, 12(2):501, 2020.

[39] C. Chittithaworn, M. A. Islam, T. Keawchana, and D. H. M. Yusuf. Factors affecting business success of small & medium enterprises (smes) in thailand. *Asian social science*, 7(5):180–190, 2011.

[40] J. Choi, B. Jeong, J. Yoon, B.-Y. Coh, and J.-M. Lee. A novel approach to evaluating the business potential of intellectual properties: A machine learning-based predictive analysis of patent lifetime. *Computers & Industrial Engineering*, 145:106544, 2020.

[41] A. C. Cooper. Challenges in predicting new firm performance. *Journal of business venturing*, 8(3):241–253, 1993.

[42] F. Corea, G. Bertinetti, and E. M. Cervellati. Hacking the venture industry: An early-stage startups investment framework for data-driven investors. *Machine Learning with Applications*, 5:100062, 2021.

[43] E. A. Cortes, M. G. Martinez, and N. G. Rubio. Multiclass corporate failure prediction by adaboost. m1. *International Advances in Economic Research*, 13(3):301–312, 2007.

[44] M. Crain. Financial markets & online advertising:reevaluating the dotcom investment bubble. *Info., comm.& society*, 17(3):371–384, 2014.

[45] Crunchbase. Emerging unicorn board companies. 2024.

[46] CSIMarket.com. Performance of the industry services. *Website*, 2023.

[47] M. Daisuke, M. Yuhei, and C. PEREZ. Forecasting firm performance with machine learning: Evidence from japanese firm-level data. Technical report, japan, 2017.

[48] D. Delen, C. Kuzey, and A. Uyar. Measuring firm performance using financial ratios: A decision tree approach. *Expert systems with applications*, 40(10):3970–3983, 2013.

[49] D. Dellermann, N. Lipusch, P. Ebel, K. M. Popp, and J. M. Leimeister. Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *arXiv preprint arXiv:2105.03360*, 2021.

[50] D. J. Denis. Entrepreneurial finance: an overview of the issues and evidence. *Journal of corporate finance*, 10(2):301–326, 2004.

[51] J. Desjardins. Visualizing 200 years of u.s. stock market sectors. *https://www.visualcapitalist.com/200-years-u-s-stock-market-sectors/, Last accessed on 2019-01-25*, 2019.

[52] S. Dibb. Market segmentation: strategies for success. *Marketing Intelligence & Planning*, 16(7):394–406, 1998.

[53] S. Edgett and S. Parkinson. The development of new financial services: identifying determinants of success and failure. *International J. of Service Industry Management*, 5(4):24–38, 1994.

[54] H. Editors. The best-performing ceos in the world 2017, 2017.

[55] H. Ernst. Patent applications and subsequent changes of performance: evidence from time-series cross-section analyses on the firm level. *Research Policy*, 30(1):143–157, 2001.

[56] S. Figini, F. Bonelli, and E. Giovannini. Solvency prediction for small and medium enterprises in banking. *Decision Support Sys.*, 102:91–97, 2017.

[57] H. Florén, J. Frishammar, V. Parida, and J. Wincent. Critical success factors in early new product development: a review and a conceptual model. *Intl. Entrepreneurship & Management J.*, 14(2):411–427, 2018.

[58] Y. Freund, R. E. Schapire, et al. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, volume 96, pages 148–156. Citeseer, 1996.

[59] Y. Fuertes-Callén, B. Cuellar-Fernández, and C. Serrano-Cinca. Predicting startup survival using first years financial statements. *J. of Small Business Management*, pages 1–37, 2020.

[60] D. Gangwani, Q. Liang, S. Wang, and X. Zhu. An empirical study of deep learning frameworks for melanoma cancer detection using transfer learning and data augmentation. In *2021 IEEE International Conference on Big Knowledge (ICBK)*, pages 38–45, 2021.

[61] D. Gangwani, X. Zhu, and B. Furht. Exploring investor-business-market interplay for business success prediction. *Journal of big Data*, 10(1):48, 2023.

[62] T. Garber, J. Goldenberg, B. Libai, and E. Muller. From density to destiny: Using spatial dimension of sales data for early prediction of new product success. *Marketing Science*, 23(3):419–428, 2004.

[63] W. Gartner, J. Starr, and S. Bhat. Predicting new venture survival: an analysis of "anatomy of a start-up." cases from inc. magazine. *Journal of Business venturing*, 14(2):215–232, 1999.

[64] 2024.

[65] P. A. Gloor, P. Dorsaz, H. Fuehres, and M. Vogel. Choosing the right friends–predicting success of startup entrepreneurs and innovators through their online social network structure. *International Journal of Organisational Design and Engineering*, 3(1):67–85, 2013.

[66] P. Gompers and J. Lerner. The determinants of corporate venture capital success: Organizational structure, incentives, and complementarities. In *Concentrated corporate ownership*, pages 17–54. University of Chicago Press, 2000.

[67] G. E. Greenley. Market orientation and company performance: empirical evidence from uk companies. *British journal of management*, 6(1):1–13, 1995.

[68] J. Greenwood, Z. Hercowitz, and P. Krusell. The role of investment-specific technological change in the business cycle. *European Economic Review*, 44(1):91–115, 2000.

[69] J. Greenwood and B. Jovanovic. Financial development, growth, and the distribution of income. *Journal of political Economy*, 98(5, Part 1):1076–1107, 1990.

[70] M. Guerzoni, C. R. Nava, and M. Nuccio. The survival of start-ups in time of crisis. a machine learning approach to measure innovation. *arXiv preprint arXiv:1911.01073*, 2019.

[71] M. Guerzoni, C. R. Nava, and M. Nuccio. Start-ups survival through a crisis. combining machine learning with econometrics to measure innovation. *Economics of Innovation and New Technology*, 30(5):468–493, 2021.

[72] B. Guo, Y. Lou, and D. Pérez-Castrillo. Investment, duration, and exit strategies for corporate and independent venture capital-backed start-ups. *Journal of Economics & Management Strategy*, 24(2):415–455, 2015.

[73] A. S. Halibas, A. S. Shaffi, and M. A. K. V. Mohamed. Application of text classification & clustering of twitter data for business analytics. In *2018 Majan international conference (MIC)*, pages 1–7. IEEE, 2018.

[74] R. S. Harris, T. Jenkinson, and S. N. Kaplan. Private equity performance: What do we know? *The Journal of Finance*, 69(5):1851–1882, 2014.

[75] R. Hawkins. Looking beyond the dot com bubble: exploring the form and function of business models in the electronic marketplace. pages 65–81, 2004.

[76] M. Heidari, S. Zad, and S. Rafatirad. Ensemble of supervised and unsupervised learning models to predict a profitable business decision. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–6. IEEE, 2021.

[77] J.-J. Hou and Y.-T. Chien. The effect of market knowledge management competence on business performance: A dynamic capabilities perspective. *Intl. J. of Electronic Business Management*, 8(2), 2010.

[78] G. Houben, W. Bakker, and P. Vergauwem. Assessing the non-financial predictors of the success and failure of young firms in the netherlands. *Economics and Applied Informatics*, (1):5–14, 2005.

[79] S. D. Hunt. A general theory of business marketing: R-a theory, alderson, the isbm framework, and the imp theoretical structure. *Industrial Marketing Management*, 42(3):283–293, 2013. Theoretical Perspectives in Industrial Marketing Management.

[80] E. Ileberi, Y. Sun, and Z. Wang. A machine learning based credit card fraud detection using the ga algorithm for feature selection. *Journal of Big Data*, 9(1):1–17, 2022.

[81] IPorg. The reachability innovation and intellectual property in business today. *https://www.wipo.int/ip-outreach/en/ipday/2017/innovation$_a$nd$_i$ntellectual$_p$roperty.html*, 2017.

[82] P. Iyer, A. Davari, M. Zolfagharian, and A. Paswan. Market orientation, positioning strategy and brand performance. *Industrial Marketing Management*, 81:16–29, 2019.

[83] A. Jaafari. Management of risks, uncertainties and opportunities on projects: time for a fundamental shift. *International Journal of Project Management*, 19(2):89–101, 2001.

[84] M. Javaid, A. Haleem, and R. P. Singh. A study on chatgpt for industry 4.0: Background, potentials, challenges, and eventualities. *Journal of Economy and Technology*, 1:127–143, 2023.

[85] S. Jhaveri, I. Khedkar, Y. Kantharia, and S. Jaswal. Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns. In *3rd Intl. Conf. on Computing Methodologies & Communication*, pages 1170–1173. IEEE, 2019.

[86] J. Jiao and Y. Zhang. Product portfolio identification based on association rule mining. *Computer-Aided Design*, 37(2):149–172, 2005.

[87] S. Jones and T. Wang. Predicting private company failure: A multi-class analysis. *J. of International Financial Markets, Institutions & Money*, 61:161–188, 2019.

[88] S. H. Jung and Y. J. Jeong. Twitter data analytical methodology development for prediction of start-up firms' social media marketing level. *Technology in Society*, 63:101409, 2020.

[89] A. Justiniano, G. E. Primiceri, and A. Tambalotti. Investment shocks and business cycles. *Journal of Monetary Economics*, 57(2):132–145, 2010.

[90] M. Kakati. Success criteria in high-tech new ventures. *Technovation*, 23(5):447–457, 2003.

[91] M. Kaur and S. Kang. Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Computer Science*, 85:78–85, 2016.

[92] K. Keasey and R. Watson. Financial distress prediction models: a review of their usefulness 1. *British J. of Management*, 2(2):89–102, 1991.

[93] W. KENTON. Fortune 100 definition, requirements, and top companies, 2022.

[94] B. Kim, H. Kim, and Y. Jeon. Critical success factors of a design startup business. *Sustainability*, 10(9):2981, 2018.

[95] J. Kim and S. Lee. Patent databases for innovation studies: A comparative analysis of uspto, epo, jpo and kipo. *Technological Forecasting & Social Change*, 92:332–345, 2015.

[96] M. Kleinaltenkamp and F. Jacob. German approaches to business-to-business marketing theory: origins and structure. *Journal of Business Research*, 55(2):149–155, 2002. Marketing Theory in the Next Millennium.

[97] W. Kluwer. Business success depends upon successful marketing, 2020.

[98] R. Kozielski. Determinants of smes business success–emerging market perspective. *International Journal of Organizational Analysis*, 27(2):322–336, 2019.

[99] A. Krishna, A. Agrawal, and A. Choudhary. Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th international conference on data mining workshops (ICDMW)*, pages 798–805. IEEE, 2016.

[100] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi. Customer churn prediction system: a machine learning approach. *Computing*, 104(2):271–294, 2022.

[101] A. Lee. Welcome to the unicorn club: Learning from billion-dollar startups. 2013.

[102] S. Lee, B. Yoon, C. Lee, and J. Park. Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technological Forecasting and Social Change*, 76(6):769–786, 2009.

[103] B. Leković and S. M. Marić. Measures of small business success/performance–importance, reliability and usability. *Industrija*, 43(2), 2015.

[104] J. Lerner, A. Schoar, S. Sokolinski, and K. Wilson. The globalization of angel investments: Evidence across countries. *Journal of Financial Economics*, 127(1):1–20, 2018.

[105] V. L. Lewis and N. C. Churchill. The five stages of small business growth. *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*, 1983.

[106] L. Leydesdorff, D. Kushnir, and I. Rafols. Interactive overlay maps for us patent (uspto) data based on international patent classification (ipc). *Scientometrics*, 98(3):1583–1599, 2014.

[107] B. Li, X. Zhu, R. Li, and C. Zhang. Rating knowledge sharing in cross-domain collaborative filtering. *IEEE Transactions on Cybernetics*, 45(5):1068–1082, 2015.

[108] J. Li. Prediction of the success of startup companies based on support vector machine and random forset. In *2020 2nd Intl. Workshop on AI and Education*, pages 5–11, 2020.

[109] Y. Li, V. Rakesh, and C. K. Reddy. Project success prediction in crowdfunding environments. In *Proceedings of 9th ACM Intl. Conf. on Web Search and Data Mining*, pages 247–256, 2016.

[110] I. Lian. Eight stages of new product.

[111] Y. E. Liang and S.-T. D. Yuan. Predicting investor funding behavior using crunchbase social network features. *Internet Research*, 2016.

[112] F. Lin, C.-C. Yeh, and M.-Y. Lee. Integrating nonlinear dimensionality reduction with random forests for financial distress prediction. *J. of Testing and Evaluation*, 43(3):645–653, 2015.

[113] W. Lin, S. A. Alvarez, and C. Ruiz. Efficient adaptive-support association rule mining for recommender systems. *Data mining & knowledge discovery*, 6(1):83–105, 2002.

[114] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai. Machine learning in financial crisis prediction: a survey. *IEEE Transactions on Systems & Cybernetics, Part C (Applications and Reviews)*, 42(4):421–436, 2011.

[115] R. E. Litan and A. M. Rivlin. Beyond the dot. coms: The economic promise of the internet, 2001.

[116] M. E. Lokanan and K. Sharma. Fraud prediction using machine learning: The case of investment advisors in canada. *Machine Learning with Applications*, 8:100269, 2022.

[117] O. Lukason and K. Käsper. Failure prediction of government funded start-up firms. *Investment Mgmt. & Financial Innovations*, 14(2):296–306, 2017.

[118] R. N. Lussier. A nonfinancial business success versus failure prediction mo. *Journal of Small Business Management*, 33(1):8, 1995.

[119] R. N. Lussier and C. E. Halabi. A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, 48(3):360–377, 2010.

[120] S. Lyu, S. Ling, K. Guo, H. Zhang, K. Zhang, S. Hong, Q. Ke, and J. Gu. Graph neural network based vc investment success prediction. *arXiv preprint arXiv:2105.11537*, 2021.

[121] I. C. MacMillan, L. Zemann, and P. Subbanarasimha. Criteria distinguishing successful from unsuccessful ventures in the venture screening process. *J. of business venturing*, 2(2):123–137, 1987.

[122] N. G. Maity and S. Das. Machine learning for improved diagnosis and prognosis in healthcare. In *2017 IEEE aerospace conference*, pages 1–9. IEEE, 2017.

[123] S. Makridakis. Factors affecting success in business: management theories/tools versus predicting changes. *European Management Journal*, 14(1):1–20, 1996.

[124] N. G. Mankiw. Real business cycles: A new keynesian perspective. *Journal of Economic Perspectives*, 3(3):79–90, September 1989.

[125] C. R. N. Marco Guerzoni and M. Nuccio. Start-ups survival through a crisis. combining machine learning with econometrics to measure innovation. *Economics of Innovation and New Technology*, 30(5):468–493, 2021.

[126] B. Marr. The 10 best examples of how companies use artificial intelligence in practice, 2019.

[127] A. Martin, M. Manjula, and D. V. P. Venkatesan. A business intelligence model to predict bankruptcy using financial domain ontology with association rule mining algorithm. *arXiv preprint arXiv:1109.1087*, 2011.

[128] L. Mbugua, P. Harris, G. Holt, and P. Olomolaiye. A framework for determining critical success factors influencing construction business performance. In *Proceedings of the Association of Researchers in Construction Management 15th Annual Conference*, volume 1, pages 255–64, 1999.

[129] N. A. Morgan. Marketing and business performance. *Journal of the academy of marketing science*, 40(1):102–119, 2012.

[130] G. A. Mousa, E. A. Elamir, and K. Hussainey. Using machine learning methods to predict financial performance: Does disclosure tone matter? *International Journal of Disclosure and Governance*, 19(1):93–112, 2022.

[131] J. V. B. Murcia. Supervised and unsupervised machine learning approaches in predicting startup success. *TWIST*, 19(1):203–208, 2024.

[132] R. Nisbet, J. Elder, and G. Miner. Handbook of statistical analysis and data mining applications, 2009.

[133] C. Pan. *Improve Entrepreneurial Funding Screening and Evaluation: Business Success Prediction with Machine Learning*. PhD thesis, Stanford University, 2021. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-01-13.

[134] C. Pan, Y. Gao, and Y. Luo. Machine learning prediction of companies' business success. *CS229: Machine Learning, Fall*, 2018.

[135] A. K. Pasayat, B. Bhowmick, and R. Roy. Factors responsible for the success of a start-up: A meta-analytic approach. *IEEE Transactions on Engineering Management*, 70(1):342–352, 2020.

[136] G. Perboli and E. Arabnezhad. A machine learning-based dss for mid and long-term company crisis prediction. *Expert Sys. with Applications*, 174:114758, 2021.

[137] A. Petropoulos, V. Siakoulis, E. Stavroulakis, and N. E. Vlachogiannakis. Predicting bank insolvencies using machine learning techniques. *International J. of Forecasting*, 36(3):1092–1113, 2020.

[138] M. R. A. Purnomo, A. Azzam, and A. U. Khasanah. Effective marketing strategy determination based on customers clustering using machine learning technique. In *Journal of Physics: Conference Series*, volume 1471, page 012023. IOP Publishing, 2020.

[139] M. Qasem, R. Thulasiram, and P. Thulasiram. Twitter sentiment classification using machine learning techniques for stock markets. In *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 834–840, 2015.

[140] S. A. Qureshi, A. S. Rehman, A. M. Qamar, A. Kamal, and A. Rehman. Telecommunication subscribers' churn prediction model using machine learning. In *Eighth international conference on digital information management (ICDIM 2013)*, pages 131–136. IEEE, 2013.

[141] A. Rai, R. Patnayakuni, and N. Patnayakuni. Technology investment and business performance. *Communications of the ACM*, 40(7):89–97, 1997.

[142] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose. Detection of financial statement fraud and feature selection using data mining techniques. *Decision support systems*, 50(2):491–500, 2011.

[143] D. Remenyi and M. Sherwood-Smith. *IT investment: making a business case*. Routledge, 2012.

[144] N. A. Richard Surya Christanto. The influence of investor demographic factors on investment decisions in the stock market with behavioral bias as a mediating variable. *THE 6th INDONESIAN FINANCE ASSOCIATION*, 2020.

[145] G. Ross, S. Das, D. Sciro, and H. Raza. Capitalvx: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, 7:94–114, 2021.

[146] N. O. Rule and N. Ambady. The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological science*, 19(2):109–111, 2008.

[147] C. Ryan. Computer and internet use in the united states, 2016. `https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-39.pdf`, Last accessed on 2018-08.

[148] T. Sampath Kumar, A. Saikiran, I. Apoorva, A. U. Kiran, M. Daddanala, and A. V. Raju. Prediction of success for currently operating startups. In *AIP Conference Proceedings*, volume 3122. AIP Publishing, 2024.

[149] P. A. Samuelson. Interactions between the multiplier analysis and the principle of acceleration. *The Review of Economics and statistics*, 21(2):75–78, 1939.

[150] I. L. Sandvik and K. Sandvik. The impact of market orientation on product innovativeness and business performance. *Intl. J. of Research in Marketing*, 20(4):355–376, 2003.

[151] J. Santisteban, D. Mauricio, and O. Cachay. Critical success factors for technology-based startups. *Intl J. of Entrepreneurship & Small Business*, 42(4):397–421, 2021.

[152] J. R. Saura, A. Reyes-Menéndez, N. deMatos, and M. B. Correia. Identifying startups business opportunities from ugc on twitter chatting: An exploratory analysis. *J. of Theoretical & Electronic Comm.*, 16(6):1929–1944, 2021.

[153] D. Schreiber-Gregory, H. Jackson, and K. Bader. Logistic and linear regression assumptions: Violation recognition and control. *Henry M Jackson Foundation*, 2018.

[154] J. A. Schumpeter. The theory of economic development: An inquiry into profits, capita i, credit, interest, and the business cycle. 2017.

[155] J. D. Šebestová and C. R. G. Popescu. Factors influencing investments into human resources to support company performance. *Journal of Risk and Financial Management*, 15(1):19, 2022.

[156] J. R. Shah and M. B. Murtaza. A neural network based clustering procedure for bankruptcy prediction. *American Business Review*, 18(2):80–86, 2000.

[157] V. Shah. Predicting the success of a startup company. *Oklahoma State University*, 2019.

[158] B. Sharchilev, M. Roizner, A. Rumyantsev, D. Ozornin, P. Serdyukov, and M. de Rijke. Web-based startup success prediction. In *Proceedings 27th ACM Intl. Conf. on Info. & Knowledge Mgmt*, pages 2283–2291, 2018.

[159] A. Sharma, D. Bhuriya, and U. Singh. Survey of stock market prediction using machine learning approach. In *2017 international conference of electronics, communication & aerospace technology (ICECA)*, volume 2, pages 506–509. IEEE, 2017.

[160] D. Singh, E. J. Leavline, S. Muthukrishnan, and R. Yuvaraj. Machine learning based business forecasting. *IJ Information Engineering and Electronic Business*, 6:40–51, 2018.

[161] E. Sivasankar, C. Selvi, and C. Mala. A study of dimensionality reduction techniques with machine learning methods for credit risk prediction. In H. S. Behera and D. P. Mohapatra, editors, *Computational Intelligence in Data Mining*, pages 65–76, Singapore, 2017. Springer Singapore.

[162] S. Y. Sohn and J. W. Kim. Decision tree-based technology credit scoring for start-up firms: Korean case. *Expert Systems with Applications*, 39(4):4007–4012, 2012.

[163] M. Solesvik and M. Gulbrandsen. Partner selection for open innovation. *Technology Innovation Management Review*, 3(4):6–11, 2013.

[164] M. Song, K. Podoynitsyna, H. Van Der Bij, and J. I. Halman. Success factors in new ventures: A meta-analysis. *Journal of product innovation management*, 25(1):7–27, 2008.

[165] Y.-g. Song, Q.-l. Cao, and C. Zhang. Towards a new approach to predict business performance using machine learning. *Cognitive Systems Research*, 52:1004–1012, 2018.

[166] G. Sonkavde, D. S. Dharrao, A. M. Bongale, S. T. Deokate, D. Doreswamy, and S. K. Bhat. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*, 11(3):94, 2023.

[167] M. Stamenković and M. M. Milanović. Outlier detection in function of quality improvement of business decisions. In *Proceedings of the International scientific conference–Enterprises in hardship: economics, managerial & juridical perspectives*, pages 173–184, 2014.

[168] R. Stuart and P. A. Abetti. Start-up ventures: Towards the prediction of initial success. *Journal of business venturing*, 2(3):215–230, 1987.

[169] J. Sun, H. Fujita, Y. Zheng, and W. Ai. Multi-class financial distress prediction based on support vector machines integrated with the decomposition and fusion methods. *Information Sciences*, 559:153–170, 2021.

[170] J. Sun, H. Li, Q.-H. Huang, and K.-Y. He. Predicting financial distress & corporate failure: A review from state-of-the-art definition, modeling, sampling & featuring approaches. *Knowledge-Based Systems*, 57:41–56, 2014.

[171] J. Szarek and J. Piecuch. The importance of startups for construction of innovative economies. *International Entrepreneurship Review*, 4(3):389, 2018.

[172] M. Takarabe, D. Shigemizu, M. Kotera, S. Goto, and M. Kanehisa. Network-based analysis and characterization of adverse drug–drug interactions. *Journal of chemical information and modeling*, 51(11):2977–2985, 2011.

[173] D. Thorleuchter, D. Van den Poel, and A. Prinzie. Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in b-to-b marketing. *Expert systems with appl.*, 39(3):2597–2605, 2012.

[174] S. Tomy and E. Pardede. From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. *Sustainability*, 10(3):602, 2018.

[175] C. T. Trapp, D. K. Kanbach, and S. Kraus. Sector coupling and business models towards sustainability: The case of the hydrogen vehicle industry. *Sustainable Technology and Entrepreneurship*, 1(2):100014, 2022.

[176] C.-F. Tsai. Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 22(2):120–127, 2009.

[177] R. Tucker. Innovation can absolutely, positively be measured. four tips from the pros for doing it right., 2021.

[178] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim. A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector. *IEEE Access*, 7:60134–60149, 2019.

[179] C. Ünal. Searching for a unicorn: A machine learning approach towards startup success prediction. Master's thesis, Humboldt-Universität zu Berlin, 2019.

[180] M. Vochozka, J. Vrbka, and P. Suler. Bankruptcy or success? the effective prediction of a company's financial development using lstm. *Sustainability*, 12(18):7529, 2020.

[181] C. S. Vui, G. K. Soon, C. K. On, R. Alfred, and P. Anthony. A review of stock market prediction with artificial neural network(ann). In *IEEE intl. conf. on control sys., comp. & eng.*, pages 477–482. IEEE, 2013.

[182] X. Wan. A literature review on the relationship between foreign direct investment and economic growth. *International Business Research*, 3(1):52, 2010.

[183] L. Wang and C. Wu. Business failure prediction based on two-stage selective ensemble with manifold learning algorithm & kernel-based fuzzy self-organizing map. *Knowledge-Based Systems*, 121:99–110, 2017.

[184] C.-P. Wei, Y.-S. Jiang, and C.-S. Yang. Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach. In *Workshop on E-Business*, pages 187–200. Springer, 2008.

[185] J. Weking, T. P. Böttcher, S. Hermes, and A. Hein. Does business model matter for startup success? a quantitative analysis. *27th European Conference on Information Systems (ECIS)*, 2019.

[186] G. Xiang, Z. Zheng, M. Wen, J. Hong, C. Rose, and C. Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *Proceedings of the international AAAI conference on web and social media*, volume 6, pages 607–610, 2012.

[187] G. Xiang, Z. Zheng, M. Wen, J. Hong, C. Rose, and C. Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1):607–610, Aug. 2021.

[188] B. Yankov, P. Ruskov, and K. Haralampiev. Models and tools for technology start-up companies success analysis. *Economic Alternatives*, 3:15–24, 2014.

[189] R. Yazdipour and R. Constand. Predicting firm failure: A behavioral finance perspective. *J. of Entrepreneurial Finance*, 14(3):90–104, 2010.

[190] B. Yeo and D. Grant. Predicting service industry performance using decision tree analysis. *Intl J. of Information Management*, 38(1):288–300, 2018.

[191] X. Yuan, F. Hou, and X. Cai. How do patent assets affect firm performance? from the perspective of industrial difference. *Technology Analysis & Strategic Management*, pages 1–14, 2020.

[192] E. L. Yuxian and S.-T. D. Yuan. Investors are social animals: Predicting investor behavior using social network features via supervised learning approach, 2013.

[193] M. Zaghloul, S. Barakat, and A. Rezk. Predicting e-commerce customer satisfaction: Traditional machine learning vs. deep learning approaches. *Journal of Retailing and Consumer Services*, 79:103865, 2024.

[194] D. Zahay and A. Griffin. Marketing strategy selection, marketing metrics, & firm performance. *J. of Business & Industrial Marketing*, 2010.

[195] M. Zane. How many new businesses started in 2022., 2023.

[196] R. Zarutskie. The role of top management team human capital in venture capital markets: Evidence from first-time funds. *Journal of Business Venturing*, 25(1):155–172, 2010.

[197] M. Zekić-Sušac, N. Šarlija, A. Has, and A. Bilandžić. Predicting company growth using logistic regression and neural networks. *Croatian operational research review*, 7(2):229–248, 2016.

[198] K. Zhang, S. Bhattacharyya, and S. Ram. Large-scale network analysis for online social brand advertising. *Mis Quarterly*, 40(4):849–868, 2016.

[199] Q. Zhang, T. Ye, M. Essaidi, S. Agarwal, V. Liu, and B. T. Loo. Predicting startup crowdfunding success through longitudinal social engagement analysis. In *Proceedings of 2017 ACM on Conf. on Information & Knowledge Management*, pages 1937–1946, 2017.

[200] S. Zhang, H. Zhong, Z. Yuan, and H. Xiong. Scalable heterogeneous graph neural networks for predicting high-potential early-stage startups. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2202–2211, 2021.

[201] X. Zhong and D. Enke. Forecasting daily stock market return using dimensionality reduction. *Expert Systems with Applications*, 67:126–139, 2017.

[202] L. Zhou and K. K. Lai. Adaboost models for corporate bankruptcy prediction with missing data. *Computational Economics*, 50(1):69–94, 2017.

[203] L. Zhou, K. K. Lai, and J. Yen. Bankruptcy prediction using svm models with a new approach to combine features selection and parameter optimisation. *International Journal of Systems Science*, 45(3):241–253, 2014.

[204] L. Zhou, K. P. Tam, and H. Fujita. Predicting the listing status of chinese listed companies with multi-class classification models. *Information Sciences*, 328:222–236, 2016.

[205] K. Żbikowski and P. Antosiuk. A machine learning, bias-free approach for predicting business success using crunchbase data. *Info. Processing Mgmt.*, 58(4):102555, 2021.