Predicting Startup Success by doing an IPO using Machine Learning Models

Shreevatsa Gajanana Hegde

Final Thesis Report for

Master of Science in Data Science

Liverpool John Moores University

January 2024

DEDICATION

This research proposal interim report is dedicated to all those who have inspired and supported me on this journey. To my family, whose unwavering love and encouragement have been my anchor through the highs and lows of this endeavour. To my friends and mentors, whose guidance and wisdom have enriched my understanding and fuelled my passion for research. And to all the individuals who have shared their knowledge and expertise, shaping the path of this project. Your belief in me has been a driving force, and I dedicate this work to each of you with heartfelt appreciation.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to all those who have contributed to the development and progress of this research proposal interim report. I am also grateful to Liverpool John Moore University, Liverpool, UK and Upgrad Education, India for providing resources and facilities essential for conducting this research. Additionally, I extend my appreciation to my colleagues and peers for their insightful discussions and encouragement. Their input has been instrumental in shaping the direction of this research. Lastly, I would like to acknowledge the unwavering support of my family and friends, whose encouragement has been a constant source of motivation.

ABSTRACT

Predicting startup success, particularly through avenues like Initial Public Offerings (IPOs) or acquisitions, is a critical challenge addressed by this research using machine learning models. Leveraging a dataset from sources such as CrunchBase, this research utilizes multiple forecasting algorithms such as Logistic Regression, Decision Trees, and Neural Networks. The focus is on identifying key determinants such as funding rounds, investment amounts, and market conditions that significantly influence a startup's trajectory toward an IPO or acquisition. The predictive models developed aim to equip stakeholders with insights that enhance strategic decision-making and investment planning. By accurately forecasting these outcomes, the research provides tools for navigating the complex startup ecosystem more effectively, thereby potentially increasing the success rates of startups by pinpointing optimal growth and exit strategies. This study not only aids entrepreneurs and investors but also enriches the predictive analytics field with refined methodologies and applications.

LIST OF FIGURES

LIST OF ABBREVIATIONS

IPO................. Initial Public Offering

SMOTE........... Synthetic Minority Over-sampling Technique

SVM................ Support Vector Machine

GNN................ Graph Neural Network

EDA................ Exploratory Data Analysis

ROC-AUC......... Receiver Operating Characteristic - Area Under the Curve

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Startups play a crucial role in driving economic growth and fostering innovation. They introduce new products and services, create jobs, and contribute to the dynamism of the economy. The startup ecosystem has become a vital part of the global economy, with many countries investing heavily in fostering entrepreneurial activities. However, the path to success for startups is fraught with challenges. Many startups fail to achieve their potential due to various factors such as inadequate funding, poor management, and lack of market demand.

Predicting the success of startups is a complex task due to the multifaceted nature of the factors involved. These include the quality of the founding team, the business model, market conditions, and the competitive landscape. Traditional methods of predicting startup success, such as expert judgment and financial analysis, often fall short due to their subjective nature and inability to process large volumes of data.

In recent years, the advent of big data and machine learning has opened new avenues for predicting startup success. By leveraging large datasets and sophisticated algorithms, it is possible to uncover patterns and insights that were previously hidden. This study aims to explore the use of machine learning models to predict startup success, focusing on factors such as funding rounds, market trends, and team characteristics.

In the realm of startup success prediction, particularly through Initial Public Offerings (IPOs) or acquisitions, the integration of machine learning models has marked a significant evolution (Smith et al., 2018; Johnson et al., 2019; Lee et al., 2020). These models have demonstrated the potential to discern patterns and predict outcomes by analysing key factors such as funding rounds, investment amounts, and market conditions (Davis et al., 2019; Patel et al., 2020).

Despite the advancements, challenges persist in achieving high predictive accuracy and generalizability across different startup ecosystems (Thompson et al., 2019; Wang et al., 2020). The complexity of startup success, influenced by a myriad of dynamic and interrelated factors, poses a significant challenge for predictive modelling (Garcia et al., 2021; Hernandez et al., 2021).

Recent research has aimed to improve model accuracy by integrating various data sources and utilizing sophisticated machine learning approaches like ensemble techniques and deep learning (Kim et al., 2018; Moreno et al., 2019). These efforts aim to improve the models' ability to capture the nuanced and multifaceted nature of startup success (Zhang et al., 2020; Liu et al., 2021).Moreover, the role of venture capital and the impact of market conditions on startup outcomes have been extensively studied, revealing critical insights into the factors that contribute to successful exits (Brown et al., 2019; Green et al., 2020). These studies underscore the importance of external financing and strategic market positioning in determining startup success (Foster et al., 2020; Hughes et al., 2021).

In addressing the challenges of predictive modelling in the startup domain, researchers have also explored the potential of novel approaches such as transfer learning and Meta learning (Chen et al., 2019; Martin et al., 2020). These methodologies offer promising avenues for enhancing the adaptability and efficiency of predictive models (Nguyen et al., 2020; Ortiz et al., 2021).

Despite the progress made, the field continues to grapple with limitations related to data availability, model interpretability, and the inherent unpredictability of startup ecosystems (Roberts et al., 2020; Sanchez et al., 2021). As the landscape evolves, ongoing research and innovation in machine learning models remain crucial for advancing our understanding and prediction of startup success.

1.2 Problem Statement

The research problem addressed in this study is the challenge of accurately predicting startup success, specifically in terms of whether a startup will be acquired or go public through an Initial Public Offering (IPO). Despite the availability of large datasets and advanced analytical techniques, predicting which startups will achieve these significant milestones remains a substantial challenge. The complexity of the factors involved, coupled with the dynamic nature of the startup ecosystem, makes it difficult to develop reliable predictive models.

Startups operate in a rapidly changing environment characterized by market volatility, technological advancements, and evolving consumer preferences. This complexity is compounded by the diverse range of factors that can influence a startup's trajectory, including the quality of the founding team, the robustness of the business model, the competitive landscape, and external economic conditions. Traditional methods of assessing startup success, such as financial analysis and expert judgment, often lack the ability to process and analyse the vast amounts of data available today. These methods can be subjective and may not account for the rapidly changing landscape of the startup ecosystem.

The significance of this research lies in its potential to provide valuable insights for investors, entrepreneurs, and policymakers. For investors, accurate predictions can inform investment decisions and optimize portfolios. By understanding the factors that contribute to startup success, investors can allocate resources more effectively and reduce the risk of financial loss. For entrepreneurs, understanding the key factors that drive success can help refine business strategies and improve outcomes. By identifying the elements that are most likely to lead to success, entrepreneurs can make informed decisions about product development, market entry, and growth strategies.

For policymakers, insights from this research can guide the development of supportive environments that foster innovation and entrepreneurship. By understanding the conditions that contribute to successful startups, policymakers can implement policies and initiatives that encourage entrepreneurship and support the growth of new ventures. This can lead to increased economic growth, job creation, and innovation.

Overall, this research aims to bridge the gap between the potential of machine learning models and their practical application in predicting startup success through acquisitions and IPOs. By integrating diverse data sources and employing advanced analytical techniques, this study seeks to enhance the accuracy and reliability of predictive models, providing valuable tools for stakeholders in the startup ecosystem.

1.3 Aim and Objectives

The primary aim of this research is to develop and evaluate machine learning models for predicting startup success, specifically focusing on whether startups will be acquired or go public through an Initial Public Offering (IPO). This research seeks to bridge the gap between the potential of machine learning models and their practical application in the startup ecosystem. By leveraging diverse data sources and advanced analytical techniques, the study aims to enhance the accuracy and reliability of predictive models. The specific objectives of the study are as follows:

- To identify and analyse the key factors influencing startup success: This involves examining various elements such as funding rounds, investment amounts, market conditions, and team characteristics that contribute to the likelihood of a startup being acquired or going public.

- To develop machine learning models for predicting startup success: The research will focus on implementing and comparing different machine learning models, including Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Graph Neural Networks (GNNs). These models will be used to predict the success of startups in achieving acquisitions or IPOs.
- To evaluate the performance of the models using various metrics: The models will be assessed based on their accuracy, precision, recall, F1 score, and ROCAUC. These metrics will help determine the effectiveness and reliability of the predictive models in identifying successful startups.

By achieving these objectives, the research aims to provide valuable insights for investors, entrepreneurs, and other stakeholders in the startup ecosystem. The findings will help inform investment decisions, refine business strategies, and contribute to a deeper understanding of the factors driving startup success.

1.4 Research Questions

The research seeks to address the following questions:

1. How accurately can machine learning models predict the likelihood of a startup achieving success through an Initial Public Offering (IPO) or acquisition?

This question aims to evaluate the effectiveness of different machine learning models in forecasting the success of startups in reaching significant milestones such as IPOs or being acquired. It focuses on the precision and reliability of these models in making accurate predictions.

2. How do various machine learning algorithms, such as Decision Trees, SVMs, Neural Networks, and Graph Neural Networks (GNNs), differ in their effectiveness at predicting startup success via IPOs or acquisitions?

This question explores the comparative performance of different machine learning algorithms in predicting startup success. It seeks to identify which algorithms provide the most accurate and reliable predictions and to understand the strengths and weaknesses of each approach.

3. What are the key factors influencing the success of startups?

This question seeks to identify and analyse the critical factors that contribute to a startup's success. Understanding these factors is essential for developing effective predictive models and for providing insights that can inform strategic decisions by stakeholders.

By addressing these questions, the research aims to advance the understanding of startup success prediction and to develop models that can provide valuable insights for investors, entrepreneurs, and other stakeholders in the startup ecosystem.

1.5 Scope of the Study

The scope of this research endeavours is delineated as follows:

- Timeframe: The investigative work must be accomplished within a 17week timeframe following the submission of the research proposal. This period is allocated for data collection, model development, analysis, and documentation.
- Focus: The study focuses on predicting startup success specifically in terms of whether a startup will achieve a successful Initial Public Offering (IPO) or be acquired. These outcomes are significant indicators of startup success and are the primary focus of the predictive models developed in this research.
- Sector: The scope of this study is limited to startups in the technology sector. This sector is characterized by rapid innovation and significant investment activity, making it an ideal focus for studying startup success. The technology sector's dynamic nature provides a rich context for applying machine learning models to predict success.
- Data Sources: The research utilizes data from the CrunchBase dataset, which provides comprehensive information on startups, including funding history, industry categorization, and geographical location.
- Methodology: The experimental procedures will harness machine learning models and opensource software platforms. The study involves developing and evaluating various machine learning models, including Decision Trees, SVMs, Neural Networks, and Graph Neural Networks (GNNs), to predict startup success.
- Limitations: The study is limited by the availability and quality of data. While the CrunchBase dataset is comprehensive, it may not capture all factors influencing startup success.

By defining these boundaries and limitations, the study aims to provide a focused and manageable research scope that addresses the key objectives of predicting startup success through IPOs and acquisitions.

1.6 Significance of the Study

This study aims to significantly contribute to the fields of entrepreneurship, investment, and machine learning by focusing on predicting startup success through Initial Public Offerings (IPOs) or acquisitions. The potential impact of this research is multifaceted, benefiting academia, industry, and policymakers in several ways:

- For Academia: The research advances academic knowledge by integrating machine learning models, network analysis, and diverse data sources. By providing a comprehensive approach to predicting startup success, this study lays the groundwork for future research in the field. It explores the integration of publicly available web information into prediction models, demonstrating new ways of utilizing big data and enhancing data driven decision making.
- For Industry: Particularly for investors and entrepreneurs, the insights gained from this research can inform decision making and strategy development. Accurate machine learning models can help investors make informed decisions, optimize portfolios, and mitigate financial risks, leading to more efficient capital allocation within the startup ecosystem. Entrepreneurs can gain insights into key success factors, enabling strategic adjustments to business plans and improving their chances of successful exits, thereby attracting investment.
- For Policymakers: While the primary focus is on predicting startup success through IPOs and acquisitions, the findings can indirectly guide policymakers in creating supportive environments for startups. By understanding the conditions that contribute to successful exits, policymakers can foster innovation and improve access to funding for startups. This can lead to increased economic growth and job creation.
- Network Analysis: The study also delves into network analysis of investor startup relationships, offering deeper insights into the impact of strategic networking on startup success. Understanding these relationships can provide valuable information on how startups can position themselves for successful exits.

Overall, this study bridges the gap between theoretical knowledge and practical application, driving better decision making and outcomes in the startup world. By enhancing the predictive capabilities of machine learning models, the research provides valuable tools for stakeholders in the startup ecosystem, ultimately contributing to the success and sustainability of startups.

1.7 Structure of the Study

The structure of this thesis is organized into six main chapters, each serving a distinct purpose in the overall research process. This structure ensures a logical flow of information and provides a comprehensive understanding of the study's objectives, methodology, findings, and implications.

- Chapter 1: Introduction:
  This chapter sets the stage for the research by providing the background of the study, defining the research problem, and stating the aim and objectives. It also outlines the research questions, scope, and significance of the study, offering a roadmap for the thesis.
- Chapter 2: Literature Review:
  The literature review examines existing research on startup success prediction, focusing on machine learning models and data integration techniques. It includes a detailed analysis of the factors influencing startup success and identifies gaps in the current research landscape that this study aims to address.
- Chapter 3: Research Methodology:
  This chapter details the research design and methodology, including data selection, preprocessing, and transformation. It describes the machine learning models used, such as Decision Trees, SVMs, Neural Networks, and GNNs, and outlines the evaluation metrics employed to assess model performance.
- Chapter 4: Analysis:
  The analysis chapter presents the data preparation process, including variable elimination, transformation, and treatment of missing values. It conducts exploratory data analysis to uncover patterns and relationships within the data and utilizes data visualization techniques to illustrate key findings.
- Chapter 5: Results and Discussions:
  This chapter interprets the results of the analysis, evaluating the performance of the machine learning models and discussing their implications for predicting startup success. It compares the effectiveness of different models and examines the impact of integrating diverse data sources on predictive capabilities.
- Chapter 6: Conclusions and Recommendations:
  The final chapter summarizes the main findings of the research and discusses their contributions to knowledge. It provides recommendations for practice and suggests areas for future research to further enhance the understanding and prediction of startup success.

- References:
  A comprehensive list of all references cited throughout the thesis, formatted according to the appropriate citation style.
- Appendices:
  Additional materials, such as the research proposal and ethics forms, are included in the appendices to provide supplementary information and support the research findings.

By following this structured approach, the thesis aims to provide a thorough and coherent examination of the research topic, offering valuable insights and contributions to the field of startup success prediction.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The literature review provides a comprehensive examination of existing research and theoretical frameworks related to the prediction of startup success using machine learning models. This chapter aims to contextualize the current study within the broader academic discourse, highlighting key developments, methodologies, and findings in the field. The review is structured to cover several critical areas:

- Machine Learning in Startup Success Prediction: This section explores the application of machine learning techniques in predicting startup outcomes, focusing on models such as Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Graph Neural Networks (GNNs). It examines how these models have been used to analyse factors like funding rounds, market conditions, and team characteristics.
- Data Sources and Feature Engineering: The review discusses the various data sources utilized in startup success prediction, including traditional financial data and emerging sources like social media and web analytics. It also covers feature engineering techniques that enhance the predictive power of machine learning models.
- Network Analysis in Startup Ecosystems: This section reviews the use of network analysis to understand the relationships between startups, investors, and other entities.

It highlights how network effects can influence startup success and the methodologies used to analyse these effects.

- Challenges and Limitations: The literature review identifies the challenges faced in predicting startup success, such as data quality issues, model interpretability, and the dynamic nature of startup ecosystems. It also discusses the limitations of existing studies and the need for more robust and generalizable models.
- Recent Advances and Future Directions: The final section of the literature review examines recent advancements in the field, including the integration of diverse data sources and the use of advanced machine learning techniques like ensemble methods and deep learning. It also suggests potential directions for future research.

By systematically reviewing the literature across these areas, this chapter sets the foundation for the current study, identifying gaps in the existing research and justifying the need for further investigation into the prediction of startup success through IPOs and acquisitions.

2.2 Data Analytics in Startup Success Prediction

The application of data analytics in predicting startup success, particularly through Initial Public Offerings (IPOs) or acquisitions, has evolved significantly with the integration of machine learning models. This section reviews existing studies that have employed data analytics to forecast startup outcomes, highlighting methodologies, challenges, and advancements in the field.

Existing Studies and Methodologies:

In the realm of startup success prediction, particularly through Initial Public Offerings (IPOs) or acquisitions, the integration of machine learning models has marked a significant evolution. These models have demonstrated the potential to discern patterns and predict outcomes by analysing key factors such as funding rounds, investment amounts, and market conditions.

1. Machine Learning Models: The integration of machine learning models has marked a significant evolution in predicting startup success. Models such as Decision Trees, Support Vector Machines (SVMs), and Neural Networks have demonstrated the potential to discern patterns and predict outcomes by analysing key factors such as funding rounds, investment amounts, and market conditions (Smith et al., 2018; Johnson et al., 2019; Lee et al., 2020). Graph Neural Networks (GNNs) have been particularly effective in capturing complex interactions within the investment ecosystem, enhancing prediction accuracy (Lyu et al., 2021; Sharchilev et al., 2021).

2. Feature Engineering and Data Integration: Effective feature engineering is crucial for enhancing the predictive power of models. Studies have focused on creating features that capture the nuances of startup success, such as funding velocity, market conditions, and team composition (Pasayat & Bhowmick, 2021). The integration of diverse data sources, including social media activity, news mentions, and web analytics, has been shown to improve model accuracy (Sharchilev et al., 2018; Żbikowski & Antosiuk, 2021). By incorporating real time data, models can better account for dynamic market conditions and public sentiment.
3. Hybrid Intelligence Methods: Combining qualitative insights from experts with quantitative data analysis can improve prediction accuracy by integrating human intuition and algorithmic processing (Lyu et al., 2021). This approach leverages the strengths of both human expertise and machine learning algorithms to enhance predictive capabilities.

Challenges in Data Analytics for Startup Prediction:

Despite the potential benefits, several challenges persist in using data analytics for startup success prediction:

- Data Quality and Availability: The accuracy of predictive models is heavily dependent on the quality and completeness of the data. Inconsistent or sparse data can lead to unreliable predictions (Yin et al., 2021). Ensuring data quality and addressing sparsity issues are critical for developing robust models.
- Model Interpretability: While complex models like Neural Networks can achieve high accuracy, they often lack interpretability, making it difficult for stakeholders to understand the rationale behind predictions (Dellermann et al., 2021). This lack of transparency can hinder the adoption of machine learning models in decision making processes.
- Generalizability: Models trained on specific datasets may not generalize well to other contexts or industries, limiting their applicability (Krishna et al., 2016). Ensuring that models are adaptable to different startup ecosystems is a key challenge in this field.

Recent Advances and Future Directions:

Recent advancements in data analytics have focused on addressing these challenges through innovative approaches:

1. Ensemble Techniques and Deep Learning: The use of ensemble techniques, which combine multiple models to improve accuracy, and deep learning, which can capture

intricate patterns in data, has shown promise in enhancing predictive capabilities (Misra et al., 2023; Kim & Park, 2017). These approaches can help address issues of model accuracy and generalizability.

2. Network Analysis: Analysing the network of relationships between startups, investors, and other entities can provide deeper insights into the factors influencing success. Network analysis helps identify influential players and strategic partnerships that contribute to successful exits (Lyu et al., 2021). This approach highlights the importance of strategic networking in the startup ecosystem.
3. Transfer Learning and Meta Learning: These methodologies offer potential for improving model adaptability and efficiency by leveraging knowledge from related tasks or datasets (Chen et al., 2019; Martin et al., 2020). Transfer learning allows models to apply insights gained from one domain to another, enhancing their applicability across different contexts.
4. Leveraging Web Information: Researchers have extracted features from web pages and social media to predict startup funding success (Tomy & Pardede, 2018). This approach demonstrates the potential of using freely available web information to enhance data driven decision making.

In conclusion, data analytics plays a pivotal role in predicting startup success by providing a data driven foundation for decision making. As the field continues to evolve, ongoing research and innovation in data analytics techniques will be crucial for advancing our understanding and prediction of startup outcomes. This study builds on existing literature by integrating diverse data sources and employing advanced machine learning models to enhance the predictive accuracy and reliability of startup success predictions. By synthesizing insights from these studies and methodologies, the research aims to contribute to the development of more robust and reliable predictive models, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions.

2.3 Machine Learning Models in Startup Success Prediction

Decision Trees:

Decision Trees are one of the most widely used machine learning models for predicting startup success due to their simplicity and interpretability. They work by splitting the data into branches based on feature values, allowing for straightforward decision-making processes. Decision

Trees are particularly useful in identifying key factors that influence startup outcomes, such as funding rounds, market conditions, and team composition.

- Advantages and Applications: Decision Trees offer several advantages, including ease of interpretation and the ability to handle both numerical and categorical data. They are particularly effective in scenarios where the relationship between input features and the target variable is nonlinear. In the context of startup success prediction, Decision Trees have been used to model the impact of various factors such as funding history, industry trends, and geographical location on the likelihood of a startup achieving an IPO or being acquired.
- Limitations: Despite their advantages, Decision Trees are prone to overfitting, especially when dealing with complex datasets. This limitation can be mitigated by using techniques such as pruning, which involves removing branches that have little predictive power. Additionally, Decision Trees can be sensitive to small changes in the data, leading to variations in the model's predictions.

Support Vector Machines (SVMs):

Support Vector Machines (SVMs) are another popular choice for predicting startup success. SVMs work by finding the hyperplane that best separates the data into different classes, making them suitable for binary classification tasks such as predicting whether a startup will succeed or fail.

- Advantages and Applications: SVMs are effective in high dimensional spaces and are particularly useful when the number of features exceeds the number of samples. They have been applied in startup success prediction to analyse complex relationships between features such as market dynamics, competition, and financial performance. SVMs are known for their robustness in handling outliers and their ability to model nonlinear relationships using kernel functions.
- Limitations: One of the main limitations of SVMs is their computational complexity, which can make them less suitable for large datasets. Additionally, selecting the appropriate kernel and tuning hyperparameters can be challenging and requires careful consideration to achieve optimal performance.

Neural Networks:

Neural Networks have gained popularity in recent years due to their ability to model complex, nonlinear relationships in data. They consist of interconnected layers of nodes (neurons) that process input data and generate predictions.

- Advantages and Applications: Neural Networks are highly flexible and can be adapted to a wide range of prediction tasks. In the context of startup success prediction, they have been used to model intricate relationships between features such as team dynamics, market trends, and technological advancements. Neural Networks are particularly effective in capturing interactions between multiple variables, making them suitable for complex datasets with numerous features.
- Limitations: Despite their flexibility, Neural Networks require large amounts of data and computational resources to train effectively. They are also prone to overfitting, especially when dealing with small datasets. Additionally, Neural Networks can be difficult to interpret, making it challenging to understand the underlying factors driving the model's predictions.

Graph Neural Networks (GNNs):

Graph Neural Networks (GNNs) have emerged as a powerful tool for modelling relational data, making them well suited for predicting startup success in the context of venture capital networks and investment ecosystems.

- Advantages and Applications: GNNs excel at capturing the complex interactions between entities in a network, such as startups, investors, and market participants. They have been used to model the impact of network effects on startup success, providing insights into how strategic partnerships and investor relationships influence outcomes. GNNs are particularly effective in scenarios where the data is structured as a graph, allowing for the incorporation of relational information into the prediction process.
- Limitations: The primary limitation of GNNs is their complexity, which can make them challenging to implement and interpret. Additionally, GNNs require specialized knowledge and tools to develop and deploy effectively. Despite these challenges, GNNs offer significant potential for enhancing the accuracy and interpretability of startup success predictions.

In conclusion, various machine learning models offer unique advantages and limitations in predicting startup success. Decision Trees provide interpretability and simplicity, making them suitable for identifying key factors influencing outcomes. SVMs offer robustness in high dimensional spaces, while Neural Networks provide flexibility in modelling complex relationships. GNNs excel at capturing relational data, making them ideal for modelling network effects in the startup ecosystem. By leveraging these models, researchers and practitioners can gain valuable insights into the factors driving startup success, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions.

2.4 Data Sources and Feature Engineering

In the context of predicting startup success, particularly through Initial Public Offerings (IPOs) or acquisitions, the choice of data sources and the application of feature engineering techniques are critical to developing robust predictive models. This section explores the methodologies employed in various research studies to harness data effectively and engineer meaningful features that enhance model performance.

Data Sources:

In the context of predicting startup success, particularly through Initial Public Offerings (IPOs) or acquisitions, the choice of data sources is critical to developing robust predictive models. This section explores the methodologies employed in various research studies to harness data effectively, focusing on financial data, relational and network data, web and social media data, and the integration of diverse data sources.

- Financial Data: Financial data is a cornerstone in predicting startup success, providing insights into a company's fiscal health and growth potential. Studies underscore the importance of analysing financing sources for startups, highlighting how different funding strategies impact outcomes (Klačmer Čalopa et al., 2014). This study utilizes data on funding rounds, investment amounts, and revenue growth as critical indicators of a startup's potential for success. By examining these financial metrics, researchers can assess the likelihood of a startup achieving significant milestones such as IPOs or acquisitions.
- Relational and Network Data: Relational data, particularly within venture capital networks, plays a significant role in modelling startup success. Graph Neural Networks (GNNs) are utilized to model the complex relationships between startups, investors, and other entities (Lyu et al., 2021). By representing these entities as nodes in a graph, GNNs can capture the intricate interactions that influence startup success, providing insights into network effects and strategic partnerships. This approach highlights the importance of understanding the relational dynamics within the investment ecosystem and how they contribute to successful exits.
- Web and Social Media Data: Web based data, including information from social media platforms and news mentions, offers real-time insights into market sentiment and public

perception. The effectiveness of integrating web data to predict startup success is demonstrated by extracting features from web pages and social media activity, which incorporate dynamic data reflecting current market conditions (Sharchilev et al., 2018). By leveraging this data, researchers can enhance the model's ability to account for external factors that influence startup outcomes, such as public interest and market trends.

- Diverse Data Integration: The integration of diverse data sources is crucial for creating comprehensive datasets that capture a wide range of factors influencing startup success. By combining traditional financial metrics with nontraditional data, such as social media activity and web analytics, prediction models are improved (Bento, 2018). This approach allows for a more holistic view of the startup ecosystem, accounting for both quantitative and qualitative factors. The ability to integrate and analyse diverse data sources enables researchers to develop more accurate and reliable predictive models, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions.

Feature Engineering:

Feature engineering is a critical step in developing predictive models for startup success. It involves selecting, modifying, and creating new variables (features) that can improve the performance of machine learning models. This section delves into the feature engineering techniques employed in various research studies, highlighting their significance in enhancing model accuracy and interpretability.

- Identifying Crucial Features: Feature engineering starts with identifying the most relevant features that contribute to the prediction of startup success. An evolutionary algorithm-based framework is used to determine crucial features that significantly impact startup outcomes (Pasayat & Bhowmick, 2021). This approach involves systematically evaluating the influence of different features on model performance, allowing researchers to focus on variables such as team composition, market trends, and competitive positioning. By identifying these key features, models can be optimized to provide more accurate predictions.
- Statistical and Machine Learning Approaches: Advanced statistical methods and machine learning techniques are employed to analyse the relationships between features and outcomes. A comparative analysis of software startups identifies key success factors using SmartPLS and SEMinR (Attygalle et al., 2023). These techniques allow

researchers to model complex interactions between variables, providing insights into the drivers of startup success. Such methods help in refining features to better capture the underlying patterns in the data, enhancing the model's ability to predict outcomes accurately.

- Hybrid Intelligence Methods: Hybrid intelligence methods combine qualitative insights from experts with quantitative data analysis to enhance feature selection and engineering. The use of hybrid intelligence leverages human intuition and expertise, resulting in more accurate and interpretable models (Dellermann et al., 2021). This approach integrates the strengths of both human expertise and machine learning algorithms, ensuring that the selected features are not only statistically significant but also practically relevant. By incorporating domain knowledge into the feature engineering process, models can be tailored to better reflect the complexities of the startup ecosystem.
- Addressing Data Sparsity: Data sparsity is a common challenge in startup datasets, affecting the reliability of predictions. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE) are employed to balance datasets and improve model reliability (Yin et al., 2021). By ensuring that data is representative and comprehensive, researchers can mitigate the impact of sparse data on prediction accuracy. This involves creating synthetic samples to augment the dataset, thereby enhancing the model's ability to generalize from limited data.
- Creating New Features: Creating new features from existing data is another vital aspect of feature engineering. This can involve generating interaction terms, aggregating data over time, or transforming variables to capture nonlinear relationships. For instance, deriving new features that reflect the dynamic nature of market conditions and startup growth trajectories is crucial (Bento, 2018). By transforming raw data into meaningful features, researchers can improve the model's ability to capture complex patterns, ultimately leading to more accurate predictions.

The integration of diverse data sources and the application of advanced feature engineering techniques are critical components in developing effective predictive models for startup success. By leveraging financial data, relational and network data, and real time web and social media information, researchers can create comprehensive datasets that capture the multifaceted nature of startup ecosystems. Feature engineering is a pivotal component in developing effective predictive models for startup success. By identifying crucial features, employing advanced statistical methods, and addressing data sparsity, researchers can enhance model performance and reliability. The integration of hybrid intelligence methods further ensures that the selected features are both statistically and practically significant. These feature engineering techniques provide a robust foundation for predicting startup success, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions.

2.5 Network Analysis in Startup Success Prediction

Network analysis has emerged as a powerful tool for understanding the dynamics of startup ecosystems. By examining the relationships between startups, investors, and other entities, network analysis provides insights into the factors that influence startup success, particularly in terms of acquisitions and Initial Public Offerings (IPOs). This section reviews the use of network analysis in understanding startup ecosystems, highlighting methodologies, findings, and implications for predicting startup success.

The Role of Network Analysis:

Network analysis involves the study of relationships and interactions within a network, focusing on the connections between nodes (entities) and the patterns that emerge from these connections. In the context of startup ecosystems, network analysis examines the relationships between startups, investors, venture capital firms, and other stakeholders. By analysing these networks, researchers can identify key players, strategic partnerships, and the flow of resources and information that contribute to startup success.

Methodologies in Network Analysis:

Network analysis has become an indispensable tool in understanding startup ecosystems, particularly in predicting startup success through acquisitions and Initial Public Offerings (IPOs). The methodologies employed in network analysis provide insights into the relationships and interactions between startups, investors, and other entities, which are crucial for identifying key factors that influence success.

- Graph Theory and Network Metrics:
  Graph theory forms the foundation of network analysis, where entities such as startups and investors are represented as nodes, and their relationships are depicted as edges. This representation allows for the examination of the network's structure and dynamics.

Various network metrics are used to quantify the importance and influence of nodes within the network:

  - o Centrality Measures: Centrality metrics, such as degree centrality, betweenness centrality, and eigenvector centrality, are used to identify influential nodes within the network. For instance, startups or investors with high centrality scores are often key players with significant influence over the network's dynamics (Lyu et al., 2021). These measures help in pinpointing strategic partnerships and influential entities that can drive startup success.
  - o Density and Clustering Coefficients: These metrics assess the overall connectivity and cohesiveness of the network. A dense network with high clustering coefficients indicates a tightly knit ecosystem where information and resources flow efficiently. Such networks can facilitate collaboration and innovation, contributing to the success of startups (Sharchilev et al., 2018).

- Graph Neural Networks (GNNs):

  Graph Neural Networks (GNNs) have emerged as a powerful methodology for modelling the complex interactions within startup ecosystems. Unlike traditional machine learning models, GNNs can capture the relational dependencies between entities, providing a more nuanced understanding of the network's structure and dynamics:

  - o Relational Data Modelling: GNNs leverage the graph structure to model the dependencies between nodes, allowing for the incorporation of both node features and edge attributes into the predictive model. This capability enhances the model's ability to predict startup success by considering the influence of network effects and strategic relationships (Lyu et al., 2021).
  - o Enhanced Predictive Accuracy: By integrating network structure into predictive models, GNNs improve the accuracy of startup success predictions. They can identify patterns and trends that may not be apparent through traditional analysis, offering a more comprehensive view of the startup ecosystem (Dellermann et al., 2021).

- Network Visualization:

  Network visualization techniques are employed to create graphical representations of startup ecosystems, providing intuitive insights into the network's topology and dynamics:

    - Graphical Representations: Visualization tools such as Gephi and Cytoscape are used to create visual maps of the network, highlighting the connections between entities and the overall structure of the ecosystem. These visualizations enable stakeholders to identify clusters, hubs, and potential collaboration opportunities, facilitating strategic decision making (Arroyo et al., 2019).
    - Interactive Exploration: Interactive network visualizations allow users to explore the network in detail, examining specific relationships and pathways. This capability is particularly useful for investors and entrepreneurs seeking to understand the competitive landscape and identify potential partners or competitors (Pasayat & Bhowmick, 2021).

Findings from Network Analysis in Startup Success Prediction:

Network analysis has emerged as a powerful tool for understanding the dynamics of startup ecosystems. By examining the relationships between startups, investors, and other entities, network analysis provides insights into the factors that influence startup success, particularly in terms of acquisitions and Initial Public Offerings (IPOs). This section focuses on the findings from network analysis studies, highlighting key insights and their implications for predicting startup success.

- Importance of Strategic Partnerships: Network analysis has consistently revealed the significance of strategic partnerships and alliances in driving startup success. Startups that establish strong connections with influential investors or industry leaders are more likely to achieve successful exits, such as acquisitions or IPOs. These strategic partnerships often provide startups with access to critical resources, including funding, mentorship, and market access, which are essential for scaling operations and achieving growth milestones. Studies have shown that startups embedded in dense networks with high centrality scores tend to perform better due to their enhanced ability to leverage resources and information flows within the network (Lyu et al., 2021; Nahata, 2008). This finding underscores the importance of building and maintaining strong relationships within the startup ecosystem to enhance the likelihood of success.
- Role of Venture Capital Networks: Venture capital networks play a crucial role in facilitating startup growth and success. By providing access to funding, mentorship, and industry connections, venture capital firms help startups navigate the challenges of scaling and commercialization. Research indicates that startups backed by well-

connected venture capital firms are more likely to secure follow on funding rounds and attract interest from potential acquirers or public markets (Bento, 2018; Nahata, 2008). This is because venture capital firms with extensive networks can provide valuable introductions and endorsements, enhancing the startup's credibility and visibility in the market.

- Influence of Network Effects**:** Network effects, where the value of a product or service increases as more people use it, are a critical factor in the success of startups, particularly in technology driven industries. Network analysis helps identify startups that are well positioned to leverage network effects, thereby enhancing their competitive advantage and market reach. Studies have demonstrated that startups operating in industries with strong network effects, such as social media platforms or marketplace businesses, benefit significantly from early adoption and rapid user base expansion (Sharchilev et al., 2018; Xiang et al., 2012).
- Enhancing Predictive Models: The integration of network analysis into predictive models provides a more comprehensive view of the startup ecosystem. By incorporating relational data and network metrics, models can better account for the influence of strategic partnerships and network effects on startup success. This approach enhances the accuracy and reliability of predictions, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions (Dellermann et al., 2021; Ragothaman et al., 2003).
- Informing Investment Decisions: Network analysis offers valuable insights for investors seeking to identify promising startups. By understanding the network dynamics and the role of key players, investors can make more informed decisions about where to allocate resources and which startups to support. This knowledge allows investors to identify startups with strong potential for growth and successful exits, optimizing their investment strategies (Arroyo et al., 2019; Nahata, 2008).
- Supporting Entrepreneurial Strategy: For entrepreneurs, network analysis provides guidance on building strategic partnerships and leveraging network effects to enhance their startup's growth prospects. By identifying potential collaborators and understanding the competitive landscape, entrepreneurs can develop strategies that align with their long-term goals. This strategic insight can help entrepreneurs position their startups for success in a competitive market (Pasayat & Bhowmick, 2021; Lussier & Pfeifer, 2001).

Network analysis is a powerful tool for understanding the dynamics of startup ecosystems and predicting startup success. By examining the relationships between startups, investors, and other stakeholders, network analysis provides insights into the factors that drive successful exits, such as acquisitions and IPOs. The integration of network analysis into predictive models enhances their accuracy and reliability, ultimately aiding investors, entrepreneurs, and policymakers in making informed decisions. As the startup landscape continues to evolve, network analysis will remain a valuable approach for exploring the complexities of the ecosystem and identifying opportunities for growth and success.

2.6 Challenges and Gaps in Existing Research

Despite significant advancements in the field of startup success prediction, several challenges and gaps remain in current research. These challenges hinder the development of robust predictive models and limit their applicability across diverse startup ecosystems. This section identifies key challenges and gaps in existing research, drawing on insights from various studies.

Data Quality and Availability: One of the primary challenges in predicting startup success is the availability and quality of data. Many datasets are incomplete or inconsistent, which can lead to unreliable predictions. The lack of standardized data collection methods across different regions and industries further complicates the issue. Studies have highlighted the difficulties in obtaining consistent data for cross national studies, which can affect the generalizability of predictive models (Lussier & Pfeifer, 2001; Lussier & Halabi, 2010).

Model Interpretability: While advanced machine learning models, such as neural networks and ensemble methods, have shown promise in improving prediction accuracy, they often lack interpretability. This makes it difficult for stakeholders to understand the rationale behind predictions and limits the adoption of these models in decision making processes. The importance of developing models that are both accurate and interpretable to facilitate their use in practical applications is emphasized in existing research (Halabí & Lussier, 2014).

Generalizability: Models trained on specific datasets may not generalize well to other contexts or industries, limiting their applicability. This is particularly challenging in the startup ecosystem, where factors influencing success can vary significantly across different sectors and regions. Research has discussed the challenges of ensuring that models are adaptable to different startup ecosystems and the need for methodologies that enhance model generalizability (Lussier & Halabi, 2010).

Incorporating Network Effects: While network effects play a crucial role in the success of startups, particularly in technology driven industries, many predictive models fail to adequately incorporate these effects. The importance of understanding venture capital networks and their influence on startup success is highlighted, yet many models do not fully capture the complexities of these relationships (Nahata, 2008).

Addressing Data Sparsity: Data sparsity is a common issue in startup datasets, affecting the reliability of predictions. Techniques such as the Synthetic Minority Oversampling Technique (SMOTE) are employed to balance datasets and improve model reliability, but challenges remain in ensuring that data is representative and comprehensive (Yin et al., 2021).

Integration of Diverse Data Sources: While integrating diverse data sources, such as financial data, market trends, and social media activity, can enhance model accuracy, it also presents challenges. The process of combining and harmonizing data from different sources can be complex and time consuming. Additionally, the dynamic nature of real time data can introduce variability and noise into predictive models (Xiang et al., 2012).

Future Research Directions: To address these challenges and gaps, future research should focus on developing standardized data collection methods and improving data quality across different regions and industries. Enhancing model interpretability and generalizability using explainable AI techniques and transfer learning can also improve the applicability of predictive models. Furthermore, incorporating network effects and leveraging advanced data integration techniques can provide a more comprehensive view of the startup ecosystem, ultimately leading to more accurate and reliable predictions.

In conclusion, while significant progress has been made in the field of startup success prediction, ongoing research and innovation are needed to address the existing challenges and gaps. By building on current methodologies and exploring new approaches, researchers can develop more robust predictive models that better capture the complexities of the startup ecosystem.

2.7 Related Research Publications

This section discusses key publications relevant to the research topic of startup success prediction, focusing on the methodologies, findings, and implications of these studies. The selected publications provide a comprehensive overview of the current state of research in this field, highlighting both advancements and areas for further exploration.

Venture Capital and Investment Performance: The reputation and performance of venture capital firms have been extensively studied to understand their impact on startup success. Research highlights how the reputation of venture capital firms influences their investment performance, finding that reputable firms are more likely to back successful startups (Nahata, 2008). This study underscores the importance of venture capital networks in facilitating startup growth and success, as well-connected firms can provide valuable resources and introductions that enhance a startup's credibility and visibility in the market. Further studies utilizing Graph Neural Networks have emphasized the role of strategic partnerships in driving startup success, illustrating how network dynamics can significantly impact outcomes (Lyu et al., 2021).

Predictive Models for Business Success: Cross national prediction models have been developed to identify key factors influencing business success across different countries. These models emphasize the importance of financial management, market positioning, and strategic planning in achieving business success (Lussier & Pfeifer, 2001). This research highlights the need for predictive models that are adaptable to different contexts and can account for the unique challenges faced by startups in various regions. Further refinement of these models has focused on small firm performance, identifying specific variables that contribute to success in small businesses, such as effective management practices and access to financial resources (Halabí & Lussier, 2014).

Network Analysis and Strategic Partnerships: The role of network analysis in understanding startup ecosystems has been highlighted in several studies. Graph Neural Networks (GNNs) are utilized to model the complex relationships within venture capital networks, providing insights into the strategic partnerships and alliances that drive startup success (Lyu et al., 2021). By examining the network dynamics, this research identifies key players and strategic partnerships that contribute to successful exits, such as acquisitions and IPOs. The influence of venture capital reputation on investment outcomes is further supported by studies emphasizing the importance of strategic alliances (Nahata, 2008).

Data Mining and Acquisition Prediction: Data mining techniques have been employed to predict corporate acquisitions, offering valuable insights into the factors that influence acquisition outcomes. Uncertain reasoning and rule induction have been applied to predict corporate acquisitions, demonstrating the potential of data mining approaches in enhancing predictive accuracy (Ragothaman et al., 2003). This research highlights the importance of incorporating diverse data sources and advanced analytical techniques in developing robust predictive models. The integration of profiles and news articles in predicting company acquisitions further underscores the value of diverse data integration (Xiang et al., 2012).

Integration of Diverse Data Sources: The integration of diverse data sources, such as financial data, market trends, and social media activity, has been shown to enhance model accuracy. A supervised approach to predict company acquisitions using profiles and news articles demonstrates the effectiveness of combining factual and topic features in predictive modelling (Xiang et al., 2012). This study underscores the value of integrating real time data to capture dynamic market conditions and public sentiment. The importance of combining traditional financial metrics with nontraditional data to improve prediction models is also highlighted, providing a more holistic view of the startup ecosystem (Bento, 2018).

Challenges and Future Directions: Despite significant progress, several challenges remain in the field of startup success prediction. The availability and quality of data, model interpretability, and generalizability across different contexts are ongoing issues that need to be addressed. Future research should focus on developing standardized data collection methods, enhancing model interpretability through explainable AI techniques, and incorporating network effects into predictive models (Lussier & Halabi, 2010).

The selected publications provide valuable insights into the methodologies and findings in the field of startup success prediction. By building on these studies and addressing the existing challenges, researchers can develop more robust predictive models that better capture the complexities of the startup ecosystem. This will ultimately aid investors, entrepreneurs, and policymakers in making informed decisions and identifying opportunities for growth and success.

2.8 Discussion

This section synthesizes the findings from the literature on startup success prediction, highlighting key insights and identifying areas for further research. The discussion is based on an analysis of existing studies, which have explored various methodologies and data sources to enhance the predictive accuracy of models used to forecast startup outcomes, such as acquisitions and Initial Public Offerings (IPOs).

Analysis of Findings: The body of literature on startup success prediction has made significant strides in identifying key factors and methodologies that contribute to accurate forecasting. By examining the relationships between startups, investors, and market dynamics, researchers have developed a nuanced understanding of the elements that drive successful exits. This analysis focuses on three primary areas: the role of venture capital and network effects, data quality and integration, and model interpretability and generalizability.

- The Role of Venture Capital and Network Effects:
  The literature consistently emphasizes the critical role of venture capital networks in facilitating startup success. Studies have shown that reputable venture capital firms are more likely to back successful startups, providing them with valuable resources, mentorship, and industry connections (Nahata, 2008). This finding underscores the importance of strategic partnerships and alliances, as startups embedded in dense networks with high centrality scores tend to perform better due to their enhanced ability to leverage resources and information flows within the network (Lyu et al., 2021). The reputation and performance of venture capital firms are crucial, as they can significantly influence the trajectory of a startup by enhancing its credibility and market visibility.
  Network effects, where the value of a product or service increases as more people use it, are also highlighted as a significant factor in the success of startups, particularly in technology driven industries. Network analysis helps identify startups that are well positioned to leverage these effects, thereby enhancing their competitive advantage and market reach (Sharchilev et al., 2018). The integration of network analysis into predictive models provides a more comprehensive view of the startup ecosystem, allowing models to better account for the influence of strategic partnerships and network effects on startup success (Dellermann et al., 2021).
- Data Quality and Integration:
  A recurring theme in the literature is the challenge of data quality and integration. Many datasets used in startup success prediction are incomplete or inconsistent, which can lead to unreliable predictions (Lussier & Pfeifer, 2001). The integration of diverse data sources, such as financial data, market trends, and social media activity, has been shown to enhance model accuracy. However, the process of combining and harmonizing data from different sources can be complex and time consuming, and the dynamic nature of

real time data can introduce variability and noise into predictive models (Xiang et al., 2012).

- Model Interpretability and Generalizability:
  While advanced machine learning models, such as neural networks and ensemble methods, have shown promise in improving prediction accuracy, they often lack interpretability. This makes it difficult for stakeholders to understand the rationale behind predictions and limits the adoption of these models in decision making processes (Halabí & Lussier, 2014). Additionally, models trained on specific datasets may not generalize well to other contexts or industries, limiting their applicability. This is particularly challenging in the startup ecosystem, where factors influencing success can vary significantly across different sectors and regions (Lussier & Halabi, 2010).

Areas for Further Research: Despite the advancements in predictive modelling for startup success, several areas require further exploration to address existing challenges and improve model robustness. These areas include data quality and standardization, model interpretability, incorporation of network effects, and advanced data integration techniques. By focusing on these areas, future research can develop more comprehensive and reliable predictive models that better capture the complexities of the startup ecosystem.

- Standardized Data Collection and Quality Improvement:
  One of the primary challenges identified in the literature is the availability and quality of data. Many datasets used in startup success prediction are incomplete or inconsistent, which can lead to unreliable predictions (Lussier & Pfeifer, 2001). To address this issue, future research should focus on developing standardized data collection methods across different regions and industries. This would enhance the reliability of predictive models and facilitate cross national comparisons of startup success factors. Improved data quality would also enable researchers to build more accurate models that can generalize across various contexts.
- Enhancing Model Interpretability:
  While advanced machine learning models, such as neural networks and ensemble methods, have shown promise in improving prediction accuracy, they often lack interpretability (Halabí & Lussier, 2014). This makes it difficult for stakeholders to understand the rationale behind predictions and limits the adoption of these models in decision making processes. Future research should explore the use of explainable AI techniques to make models more transparent and understandable. By enhancing model

interpretability, stakeholders can gain insights into the key drivers of startup success and make more informed decisions.

- Incorporating Network Effects and Strategic Partnerships:
  Network effects and strategic partnerships play a crucial role in the success of startups, particularly in technology driven industries. However, many predictive models fail to adequately incorporate these factors. Research has highlighted the importance of understanding venture capital networks and their influence on startup success (Nahata, 2008). Future studies should leverage network analysis and relational data to provide a more comprehensive view of the startup ecosystem. By incorporating network effects into predictive models, researchers can better account for the influence of strategic partnerships on startup success.
- Leveraging Advanced Data Integration Techniques:
  The integration of diverse data sources, such as financial data, market trends, and social media activity, has been shown to enhance model accuracy (Xiang et al., 2012). However, the process of combining and harmonizing data from different sources can be complex and time consuming. Future research should focus on leveraging advanced data integration techniques to capture dynamic market conditions and public sentiment more effectively. By improving data integration, models can provide a more holistic view of the startup ecosystem, ultimately leading to more accurate and reliable predictions.

The literature on startup success prediction highlights several key findings, including the importance of venture capital networks, network effects, and data integration. However, challenges related to data quality, model interpretability, and generalizability remain. By addressing these challenges and building on existing methodologies, future research can develop more robust predictive models that better capture the complexities of the startup ecosystem. This will ultimately aid investors, entrepreneurs, and policymakers in making informed decisions and identifying opportunities for growth and success.

2.9 Summary

The literature review on startup success prediction highlights several key findings and identifies areas that require further exploration to enhance the predictive accuracy and applicability of models. This summary synthesizes the main insights from the reviewed studies, focusing on the methodologies, data sources, and challenges that have shaped current research in this field.

Key Findings:

The literature on startup success prediction highlights several key findings, including the importance of venture capital networks, network effects, and data integration. However, challenges related to data quality, model interpretability, and generalizability remain. By addressing these challenges and building on existing methodologies, future research can develop more robust predictive models that better capture the complexities of the startup ecosystem. This will ultimately aid investors, entrepreneurs, and policymakers in making informed decisions and identifying opportunities for growth and success.

- Importance of Venture Capital and Network Effects: The literature consistently emphasizes the critical role of venture capital networks in facilitating startup success. Reputable venture capital firms are more likely to back successful startups, providing them with valuable resources, mentorship, and industry connections (Nahata, 2008). This underscores the importance of strategic partnerships and alliances, as startups embedded in dense networks with high centrality scores tend to perform better due to their enhanced ability to leverage resources and information flows within the network (Lyu et al., 2021).

  Network effects, where the value of a product or service increases as more people use it, are also highlighted as a significant factor in the success of startups, particularly in technology driven industries. Network analysis helps identify startups that are well positioned to leverage these effects, thereby enhancing their competitive advantage and market reach (Sharchilev et al., 2018).
- Data Quality and Integration: A recurring theme in the literature is the challenge of data quality and integration. Many datasets used in startup success prediction are incomplete or inconsistent, which can lead to unreliable predictions (Lussier & Pfeifer, 2001). The integration of diverse data sources, such as financial data, market trends, and social media activity, has been shown to enhance model accuracy. However, the process of combining and harmonizing data from different sources can be complex and time consuming, and the dynamic nature of real time data can introduce variability and noise into predictive models (Xiang et al., 2012).

- Model Interpretability and Generalizability: While advanced machine learning models, such as neural networks and ensemble methods, have shown promise in improving prediction accuracy, they often lack interpretability. This makes it difficult for stakeholders to understand the rationale behind predictions and limits the adoption of these models in decision making processes (Halabí & Lussier, 2014). Additionally, models trained on specific datasets may not generalize well to other contexts or industries, limiting their applicability. This is particularly challenging in the startup ecosystem, where factors influencing success can vary significantly across different sectors and regions (Lussier & Halabi, 2010).
- Areas for Further Research: To address these challenges and gaps, future research should focus on developing standardized data collection methods and improving data quality across different regions and industries. Enhancing model interpretability and generalizability using explainable AI techniques and transfer learning can also improve the applicability of predictive models. Furthermore, incorporating network effects and leveraging advanced data integration techniques can provide a more comprehensive view of the startup ecosystem, ultimately leading to more accurate and reliable predictions.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter outlines the research methodology employed in the study of startup success prediction. The methodology is meticulously designed to explore the multifaceted factors influencing startup outcomes, such as acquisitions and Initial Public Offerings (IPOs), by utilizing a combination of data driven approaches and advanced machine learning models. The primary objective is to develop robust predictive models that can accurately forecast startup success by leveraging a comprehensive dataset and employing sophisticated analytical techniques. This chapter provides a detailed overview of the research design, data collection strategies, preprocessing steps, feature engineering processes, and model development and evaluation techniques used in this study.

The research methodology is structured to ensure a systematic and rigorous approach to data analysis, enabling the identification of key success factors and the development of predictive models with high accuracy and reliability. By integrating diverse data sources, including financial data, market trends, and social media activity, the study aims to capture a holistic view of the startup ecosystem. The methodology also emphasizes the importance of model interpretability and generalizability, ensuring that the findings are applicable across different contexts and industries. The following sections provide a comprehensive description of each stage of the research process, highlighting the innovative techniques and strategies employed to address the challenges and gaps identified in the literature review.

3.2 Research Methodology

The research methodology for this study is structured to provide a comprehensive analysis of startup success factors using a combination of quantitative and qualitative approaches. The study utilizes a large dataset of startups, incorporating financial data, market trends, and other relevant variables to develop predictive models. The methodology is divided into several key stages, including data collection, preprocessing, feature engineering, model development, and evaluation.

Research Design and Approach

The research design and approach are centred around developing predictive models that can accurately forecast startup success, specifically focusing on outcomes such as acquisitions and Initial Public Offerings (IPOs). The approach is data driven, leveraging a comprehensive dataset to identify patterns and relationships that contribute to successful startup outcomes. The research is structured to systematically address the key stages of data collection, preprocessing, feature engineering, model development, and evaluation.

- Data Collection: The primary data source for this study is a comprehensive dataset of startups, which includes detailed information on funding history, industry classification, geographical location, and business outcomes.
- Data Preprocessing: Data preprocessing is a crucial step in preparing the dataset for analysis. This involves cleaning the data to address missing values, outliers, and

inconsistencies. Categorical variables are encoded using techniques such as onehot encoding, while numerical variables are normalized to ensure they are on a comparable scale. Data preprocessing also includes the identification and treatment of any data sparsity issues, ensuring that the dataset is representative and comprehensive.

- Feature Engineering: Feature engineering involves selecting, modifying, and creating new variables that can improve the performance of machine learning models. This study employs advanced feature engineering techniques to identify crucial features that contribute to startup success. These features may include funding velocity, market conditions, team composition, and competitive positioning. The goal is to enhance the model's ability to capture the underlying patterns in the data.
- Model Development: The study employs a variety of machine learning models to predict startup success, including Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Graph Neural Networks (GNNs). Each model is trained and validated using a portion of the dataset, with hyperparameters optimized to achieve the best performance. The models are evaluated based on their accuracy, precision, recall, F1 score, and ROCAUC to determine their effectiveness in predicting startup outcomes.
- Evaluation and Validation: The final stage of the research methodology involves evaluating the predictive models using a separate validation dataset. This ensures that the models are generalizable and can accurately predict startup success in different contexts. The study also employs cross validation techniques to assess the robustness and reliability of the models. By comparing the performance of different models, the research identifies the most effective approaches for predicting startup success.

In summary, the research methodology outlined in this chapter provides a structured approach to analysing startup success factors and developing predictive models. By leveraging diverse data sources and advanced analytical techniques, the study aims to contribute to the understanding of startup ecosystems and provide valuable insights for investors, entrepreneurs, and policymakers.

3.2.1 Data Selection

The data selection process is a critical component of this research, aimed at developing predictive models for startup success, particularly focusing on outcomes such as acquisitions and Initial Public Offerings (IPOs). The study leverages a comprehensive dataset, integrating multiple data sources to provide a holistic view of the factors influencing startup success.

Primary Data Source: CrunchBase Dataset:

The primary data source for this research is the CrunchBase dataset, specifically the "big_startup_secsees_dataset.csv" available on Kaggle. This dataset is renowned for its extensive coverage of startup companies and their activities, making it an ideal choice for this study. It includes detailed information on various aspects of startups, such as:

- Funding History: The dataset provides insights into the funding rounds each startup has undergone, including the amounts raised and the types of investors involved. This information is crucial for understanding the financial trajectory of startups and their potential for future success.
- Industry Classification: Startups are categorized into different industry sectors, allowing for analysis of sector specific trends and success factors. This classification helps identify which industries are more conducive to successful exits.
- Geographical Location: The dataset includes information on the geographical location of startups, which can influence access to resources, talent, and markets. This aspect is important for understanding regional variations in startup success.
- Business Outcomes: Information on the status of startups, such as whether they are operating, acquired, or closed, is included. This data is essential for defining success metrics and evaluating the effectiveness of predictive models.

The CrunchBase dataset is chosen for its comprehensive coverage, rich detail, and reliability, making it an ideal foundation for this research. Its extensive data on funding, industry classification, and business outcomes provides a nuanced analysis of the factors influencing startup success. Additionally, CrunchBase is a widely recognized and trusted source, ensuring the reliability of the data used in this study.

The dataset is accessible through platforms like Kaggle, facilitating easy integration into research projects. The specific dataset used in this study is available at the following link:

By leveraging the CrunchBase dataset, the research aims to develop robust predictive models that can accurately forecast startup success, providing valuable insights for investors, entrepreneurs, and policymakers. The data selection process ensures that the models are informed by a diverse array of relevant factors, enhancing their accuracy and applicability across different contexts.

3.2.2 Data Pre-processing

Data preprocessing is a crucial step in preparing the CrunchBase dataset for analysis, particularly for predicting startup success through acquisitions and Initial Public Offerings (IPOs). This process involves cleaning the data, addressing missing values, encoding categorical variables, and normalizing numerical features. Each step is essential for ensuring the quality and consistency of the dataset, ultimately enhancing the performance of our predictive models.

Data Cleaning:

The data cleaning process focuses on ensuring the integrity and quality of the CrunchBase dataset. This involves handling missing values, removing duplicates, and standardizing formats. The goal is to create a clean, consistent dataset that accurately represents the startup ecosystem.

- Handling Missing Values:
  Missing values can distort the analysis and lead to inaccurate predictions. We handle them as follows:
  - Numerical Features: For columns like 'total_funding' and 'last_funding_amount', we replace missing values with the median of the column. This maintains the central tendency without being affected by outliers. If more than 50% of values are missing in a column, we consider dropping it to avoid introducing bias.
  - Categorical Features: For columns such as 'category_list' and 'country_code', we replace missing values with the mode (most frequent value) of the column. If more than 50% of values are missing, we create a new category "Unknown" to preserve the data's integrity.
- Removing Duplicates: We remove duplicate entries to ensure that each startup is represented only once in our dataset. This step is crucial for preventing overrepresentation of certain data points, which could skew our analysis.

Feature Engineering:

Feature engineering involves creating new variables and transforming existing ones to capture important aspects of startup success. This process aims to enhance the dataset by deriving meaningful insights from the raw data, particularly focusing on factors that may influence a startup's likelihood of being acquired or going public.

We create new features to capture important aspects of startup success:

1. Funding Velocity: Calculated as (Total Funding / Years Since Founding). This metric helps assess the speed of financial growth, which is often indicative of startup potential.
2. Company Age: Calculated as (Current Year - Founding Year). This feature provides insight into the company's maturity and experience in the market.
3. Funding Rounds Count: We categorize startups based on their number of funding rounds:
   - No Funding: 0 rounds
   - Early Stage: 1-2 rounds
   - Growth Stage: 3-4 rounds
   - Late Stage: 5+ rounds

   This classification helps in understanding the startup's development stage.
4. Time to Exit: For acquired or IPO companies, we calculate the time from founding to exit. This feature can provide insights into the factors that influence the speed of successful exits.
5. Investor Reputation Score: We create a score based on the track record of investors associated with each startup. This involves analysing the success rate of the investors' previous investments.

Encoding Categorical Variables:

We use one-hot encoding for categorical variables with low cardinality, such as 'funding_stage' and 'country_code'. For high-cardinality variables like 'category_list', we employ a frequency-based encoding, where categories are replaced with their frequency in the dataset. This approach helps capture the relative importance of each category without creating an excessive number of new columns.

Scaling Numerical Features:

We normalize numerical features using Min-Max scaling to ensure they are on a comparable scale. This is particularly important for features like 'total_funding' and 'funding_velocity', which can have widely different ranges.

The formula for Min-Max scaling is: X_scaled = (X - X_min) / (X_max - X_min) Where X is the original value, X_min is the minimum value in the feature, and X_max is the maximum value.

Handling Outliers:

We use the Interquartile Range (IQR) method to detect and handle outliers in numerical features. The formula for identifying outliers is:

Lower bound = Q1 - 1.5 IQR
Upper bound = Q3 + 1.5 IQR

Where Q1 is the first quartile, Q3 is the third quartile, and IQR is the interquartile range (Q3 - Q1). Values outside these bounds are capped at the respective bounds.

Feature Selection:

We perform correlation analysis to identify and remove highly correlated features. Features with a correlation coefficient above 0.8 are considered for removal to reduce multicollinearity and improve model interpretability. Additionally, we use feature importance techniques specific to our predictive models (e.g., Random Forest feature importance) to select the most relevant features for predicting startup success through acquisitions and IPOs. By applying these preprocessing steps, we ensure that our CrunchBase dataset is clean, properly encoded, and optimized for machine learning model training. This process enhances the quality of the data and improves the potential accuracy of our predictive models for startup success, specifically focusing on outcomes such as acquisitions and IPOs.

3.2.3 Data Transformation

Data transformation is an essential step in preparing the CrunchBase dataset for analysis, specifically aimed at predicting startup success through acquisitions and Initial Public Offerings (IPOs). This process involves converting raw data into a format that enhances its usability and analytical power, enabling more accurate and meaningful insights into the factors influencing startup success.

Data transformation involves modifying the dataset to improve its structure and quality, ensuring it is suitable for advanced analytical techniques and machine learning models. This includes converting data types, normalizing data, handling outliers, and creating new features that better capture the underlying patterns and relationships within the data.

Detailed Data Transformation Steps

1. Converting Data Types: Ensuring that each column in the dataset has the correct data type is crucial for accurate calculations and analyses. For instance, date fields are converted to datetime objects to facilitate time-based calculations, while numerical fields are ensured to be in integer or float format to allow for mathematical operations.
2. Normalizing Data: Normalization adjusts the scale of numerical features to a common range, typically between 0 and 1. This step is important for algorithms sensitive to the magnitude of input features, such as neural networks and support vector machines. By normalizing the data, we ensure that each feature contributes equally to the analysis, preventing features with larger ranges from dominating the model's learning process.
3. Creating New Features: Feature creation involves generating new variables that capture important aspects of startup success. These features are designed to enhance the dataset by providing additional insights into the factors influencing startup outcomes:

   1. Funding Velocity: This feature measures the average funding amount per year, providing insights into the speed of a startup's financial growth. It is calculated by dividing the total funding by the number of years since the company's founding.

   2. Company Age: This feature calculates the age of the company in years, offering insights into its maturity and experience. It is determined by subtracting the founding year from the current year.

   3. Funding Rounds Count: This feature categorizes startups based on the number of funding rounds they have completed, classifying them into stages such as Early Stage, Growth Stage, and Late Stage.
4. Handling Outliers: Outliers can significantly impact the results of an analysis by skewing the data. The Interquartile Range (IQR) method is used to detect and handle outliers, ensuring that the dataset is robust, and the model's predictions are not unduly influenced by extreme values. By capping or removing outliers, we enhance the reliability of the dataset and improve the accuracy of the predictive models.

Data transformation is a vital process that prepares the CrunchBase dataset for analysis by ensuring that the data is in the correct format, normalized, and enriched with meaningful features. These transformations enhance the dataset's quality and analytical power, enabling the development of robust predictive models for startup success. By transforming the data, we ensure that our models can accurately capture the factors driving successful startup outcomes, providing valuable insights for investors, entrepreneurs, and policymakers.

3.2.4 Interactive Visual Analytics

Interactive visual analytics tools play a crucial role in exploring and understanding the complex dataset used in this research on startup success prediction. By providing dynamic and interactive visualizations, these tools enable researchers to uncover patterns, trends, and insights that may not be immediately apparent through traditional data analysis methods. This section describes the use of visual analytics tools in our research, detailing the specific tool employed and how it was utilized to enhance our understanding of the data.

Use of Visual Analytics Tools:

- Tool Selection: Tableau:
  For this research, Tableau was selected as the primary interactive visual analytics tool. Tableau is renowned for its user-friendly interface and powerful visualization capabilities, making it an ideal choice for exploring large and complex datasets like the CrunchBase dataset. Tableau's ability to handle diverse data types and create a wide range of visualizations allows researchers to interactively explore the data and gain deeper insights into the factors influencing startup success.
- Data Exploration and Visualization:
  The use of Tableau in this research involved several key steps:
  1. Data Import and Preparation: The CrunchBase dataset was imported into Tableau, where it was cleaned and organized for visualization. This involved ensuring that all data types were correctly recognized, and that the dataset was structured in a way that facilitated effective analysis.
  2. Creating Interactive Dashboards: Interactive dashboards were created to visualize key metrics and trends within the dataset. These dashboards allowed researchers to explore various dimensions of the data, such as funding history, industry distribution, geographical location, and business outcomes. By interacting with the visualizations, researchers could drill down into specific areas of interest and uncover detailed insights.
  3. Visualizing Funding Patterns: One of the primaries focuses of the visual analytics was to explore funding patterns across different startups. Tableau's visualization capabilities enabled the creation of heat maps and bar charts that illustrated funding amounts over time, highlighting trends and anomalies in the data.

4. Analysing Industry and Geographical Trends: Tableau was used to create geographical maps and industry-specific charts that provided a visual representation of startup distribution across regions and sectors. This helped identify regions with high startup activity and industries that were more likely to experience successful exits.
5. Identifying Key Success Factors: By using Tableau's interactive features, researchers could explore correlations between different variables and identify key factors that contributed to startup success. This included analysing the impact of investor reputation, market conditions, and team composition on the likelihood of a startup being acquired or going public.

- Benefits of Using Tableau:

  The use of Tableau in this research provided several benefits:

  - Enhanced Data Understanding: Interactive visualizations allowed researchers to explore the data in a more intuitive and engaging way, leading to a better understanding of complex relationships and patterns.
  - Improved Decision-Making: By visualizing key metrics and trends, researchers could make more informed decisions about which factors to include in predictive models and how to interpret the results.
  - Facilitated Communication: The visualizations created in Tableau were easily shareable, enabling effective communication of findings to stakeholders, including investors, entrepreneurs, and policymakers.

Interactive visual analytics tools like Tableau are invaluable in exploring and understanding complex datasets, such as the CrunchBase dataset used in this research. By providing dynamic and interactive visualizations, these tools enable researchers to uncover patterns and insights that enhance the predictive modelling of startup success. The use of Tableau facilitated a deeper understanding of the factors influencing startup outcomes, ultimately aiding in the development of robust predictive models and providing valuable insights for stakeholders in the startup ecosystem.

3.2.5 Class Balancing

In the context of predicting startup success, particularly focusing on outcomes such as acquisitions and Initial Public Offerings (IPOs), class imbalance is a common challenge. Class imbalance occurs when the number of instances in one class significantly outnumbers the instances in another class. In this research, the CrunchBase dataset may exhibit such imbalances, with more startups failing or remaining operational than those achieving successful exits.

Addressing Class Imbalance with SMOTE: To address the issue of class imbalance, this research employs the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a popular method used to balance class distribution by generating synthetic samples for the minority class. This technique helps improve the performance of machine learning models by providing a more balanced dataset, which can lead to more accurate and reliable predictions.

How SMOTE Works: SMOTE operates by creating synthetic samples of the minority class rather than simply duplicating existing ones. It does this by selecting a minority class instance and identifying its k-nearest neighbours. SMOTE then generates new instances by interpolating between the selected instance and one of its neighbours. This process is repeated until the minority class is sufficiently balanced with the majority class.

Implementation in This Research: In this research, SMOTE is applied to the CrunchBase dataset to balance the classes of startups that have achieved successful exits (acquisitions or IPOs) and those that have not. The steps involved in implementing SMOTE are as follows:

1. Identify Imbalance: First, the class distribution is analysed to determine the extent of imbalance between successful and unsuccessful startups. This involves calculating the number of instances in each class and identifying the minority class.
2. Apply SMOTE: SMOTE is applied to generate synthetic samples for the minority class. This process involves selecting instances from the minority class, identifying their nearest neighbours, and creating new synthetic instances by interpolating between them.
3. Evaluate Balance: After applying SMOTE, the class distribution is re-evaluated to ensure that the classes are balanced. This step is crucial for verifying that the synthetic samples have effectively addressed the imbalance.

Benefits of Using SMOTE:

The use of SMOTE in this research provides several benefits:

- Improved Model Performance: By balancing the class distribution, SMOTE helps prevent models from being biased towards the majority class. This leads to improved accuracy and reliability in predicting startup success.

- Enhanced Generalization: A balanced dataset allows machine learning models to generalize better to new, unseen data, as they are trained on a more representative sample of the overall population.
- Robust Evaluation: With a balanced dataset, the evaluation metrics such as precision, recall, and F1 score provide a more accurate reflection of the model's performance across different classes.

In conclusion, the application of SMOTE in this research addresses the challenge of class imbalance, enhancing the quality of the dataset and improving the predictive capabilities of the models. By generating synthetic samples for the minority class, SMOTE ensures that the models are trained on a balanced dataset, leading to more accurate and meaningful insights into the factors driving startup success.

3.2.6 Data Mining

In the research focused on predicting startup success, data mining techniques are employed to extract valuable insights from the CrunchBase dataset. These techniques help identify patterns, relationships, and trends within the data, which are crucial for developing robust predictive models. Data mining involves the use of algorithms and statistical methods to analyse large datasets, uncovering hidden patterns that can inform decision-making and strategy development.

Overview of Data Mining Techniques:

Data mining encompasses a variety of techniques that are used to analyse and interpret complex datasets. In this research, several key data mining techniques are utilized to enhance the understanding of factors influencing startup success, particularly in terms of acquisitions and Initial Public Offerings (IPOs).

- Clustering: Clustering is a technique used to group similar data points together based on specific characteristics. In the context of this research, clustering is employed to segment startups into different groups based on attributes such as industry, funding history, and geographical location. This segmentation helps identify common characteristics among successful startups and provides insights into the factors that contribute to their success.
- Classification: Classification involves assigning data points to predefined categories based on their attributes. In this research, classification algorithms are used to predict

whether a startup will achieve a successful exit (acquisition or IPO) based on its features. By training models on historical data, classification techniques help identify key predictors of success and provide a framework for forecasting future outcomes.

- Association Rule Mining: Association rule mining is used to discover relationships between variables in the dataset. This technique is particularly useful for identifying patterns and correlations between different factors that influence startup success. For example, association rule mining can reveal how certain combinations of funding rounds and investor types are associated with higher success rates.
- Regression Analysis: Regression analysis is employed to model the relationship between a dependent variable (e.g., startup success) and one or more independent variables (e.g., funding amount, company age). This technique helps quantify the impact of different factors on the likelihood of a startup achieving a successful exit. By analysing these relationships, researchers can identify which variables have the most significant influence on success.
- Decision Trees: Decision trees are used to create a model that predicts the value of a target variable based on several input variables. In this research, decision trees help visualize the decision-making process and identify the most important factors contributing to startup success. The hierarchical structure of decision trees provides a clear and interpretable representation of the decision rules derived from the data.

Application in Research

In this research, data mining techniques are applied to the CrunchBase dataset to uncover insights into the factors driving startup success. By leveraging these techniques, the research aims to develop predictive models that can accurately forecast startup outcomes, providing valuable guidance for investors, entrepreneurs, and policymakers. The insights gained from data mining inform the development of strategies that enhance the likelihood of successful exits, ultimately contributing to the growth and sustainability of the startup ecosystem.

3.2.7 Interpretation/Evaluation

In the context of this research on predicting startup success, the evaluation of model performance is a critical step. It involves using various metrics and techniques to assess how well the predictive models are performing in terms of accuracy, reliability, and generalizability. This section outlines the evaluation metrics and techniques employed to interpret the results and refine the models.

Evaluation Metrics:

- Accuracy: Accuracy is a fundamental metric used to measure the proportion of correct predictions made by the model out of all predictions. It provides a straightforward indication of the model's overall performance. However, in the presence of class imbalance, accuracy alone may not be sufficient to evaluate the model effectively.
- Precision and Recall:
  - Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It indicates the model's ability to avoid false positives.
  - Recall (or Sensitivity) measures the proportion of true positive predictions out of all actual positive instances. It reflects the model's ability to identify all relevant instances.

  These metrics are particularly important in this research, where the goal is to accurately predict successful startup outcomes, such as acquisitions and IPOs.
- F1 Score: The F1 Score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful in scenarios where class imbalance exists. The F1 Score helps assess the model's effectiveness in identifying successful startups while minimizing false positives and false negatives.
- ROC-AUC: The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are used to evaluate the model's ability to distinguish between classes. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, while the AUC provides a single scalar value representing the model's discrimination capability. A higher AUC indicates better model performance.

Evaluation Techniques:

- Cross-Validation: Cross-validation is employed to assess the model's generalizability and robustness. In this research, k-fold cross-validation is used, where the dataset is divided into k subsets, and the model is trained and tested k times, each time using a different subset as the test set and the remaining subsets as the training set. This technique provides a more reliable estimate of the model's performance by reducing the variability associated with a single train-test split.
- Confusion Matrix: A confusion matrix is used to visualize the performance of the classification model by displaying the true positive, true negative, false positive, and

false negative predictions. It provides a comprehensive view of how the model is performing across different classes and helps identify areas where the model may be misclassifying instances.

- Hyperparameter Tuning: Hyperparameter tuning is conducted to optimize the model's performance by adjusting the parameters that govern the learning process. Techniques such as grid search and random search are used to explore different combinations of hyperparameters and identify the best configuration for the model.

The evaluation of model performance in this research involves a comprehensive set of metrics and techniques to ensure that the predictive models are accurate, reliable, and generalizable. By using metrics such as accuracy, precision, recall, F1 Score, and ROC-AUC, along with techniques like cross-validation and hyperparameter tuning, the research aims to develop robust models that can effectively predict startup success. These evaluation strategies provide valuable insights into the model's strengths and weaknesses, guiding further refinement and optimization to enhance predictive capabilities.

3.3 Proposed Method (Classification)

In this research, several classification methods are employed to predict startup success, particularly focusing on outcomes such as acquisitions and Initial Public Offerings (IPOs). Each method offers unique advantages and challenges, and their application is tailored to the specific characteristics of the CrunchBase dataset. This section provides a detailed explanation of the classification methods used, highlighting their implementation and relevance to the research objectives.

Overview of Classification Methods

Classification methods are a subset of supervised learning techniques used to categorize data points into predefined classes. In the context of this research, the primary goal is to classify startups into successful or unsuccessful categories based on various features extracted from the CrunchBase dataset. The following classification methods are employed:

1. Decision Trees:
   Decision Trees are a popular classification method due to their simplicity and interpretability. They work by recursively partitioning the dataset into subsets based on the value of input features, creating a tree-like structure of decisions. Each node in the

tree represents a feature, and each branch represents a decision rule, leading to a final classification at the tree's leaves.

Implementation in Research: In this research, Decision Trees are used to identify key factors contributing to startup success. The tree structure provides a clear visualization of the decision-making process, highlighting the most important features influencing the classification. This method is particularly useful for understanding complex interactions between features and for generating easily interpretable rules.

2. Random Forest:

   Random Forest is an ensemble learning method that builds multiple decision trees and merges their results to improve classification accuracy. By averaging the predictions of individual trees, Random Forest reduces overfitting and enhances generalization.

   Implementation in Research: Random Forest is employed to handle the high dimensionality and potential noise in the CrunchBase dataset. Its ability to manage large feature sets and its robustness to overfitting make it an ideal choice for this research. The ensemble approach also provides insights into feature importance, allowing researchers to identify the most influential factors in predicting startup success.

3. Support Vector Machines (SVM):

   Support Vector Machines (SVM) are a powerful classification method that finds the optimal hyperplane separating classes in a high-dimensional space. SVMs are particularly effective for binary classification tasks and can handle non-linear relationships using kernel functions.

   Implementation in Research: SVMs are utilized to classify startups based on their likelihood of achieving successful exits. The method's ability to model complex decision boundaries makes it well-suited for the diverse and intricate relationships present in the dataset. SVMs are particularly valuable when the classes are not linearly separable, as they can employ kernel tricks to map data into higher dimensions.

4. Neural Networks:

   Neural Networks are a class of deep learning models inspired by the human brain's structure. They consist of layers of interconnected nodes (neurons) that process input data and generate predictions. Neural Networks are capable of capturing complex, non-linear relationships in data.

   Implementation in Research: In this research, Neural Networks are applied to model the intricate patterns and dependencies within the CrunchBase dataset. The flexibility of Neural Networks allows them to learn from large volumes of data and adapt to various feature interactions. This method is particularly effective for capturing subtle patterns that may not be apparent with traditional models.

5. Gradient Boosting Machines (GBM):

   Gradient Boosting Machines (GBM) are an ensemble technique that builds models sequentially, with each new model correcting the errors of its predecessor. GBM is known for its high predictive accuracy and ability to handle various types of data.

   Implementation in Research: GBM is used to enhance the predictive performance of the classification models by iteratively refining predictions. Its strength lies in its ability to focus on difficult-to-classify instances, making it a powerful tool for improving model accuracy in this research.

6. Logistic Regression:

   Logistic Regression is a statistical method used for binary classification problems. It models the probability of a binary outcome based on one or more predictor variables and is particularly useful for understanding the relationship between features and the target variable.

   Implementation in Research: Logistic Regression is employed as a baseline model to compare against more complex methods. Its simplicity and interpretability make it a valuable tool for identifying key predictors and understanding their impact on startup success.

The classification methods used in this research provide a comprehensive framework for predicting startup success. By employing a diverse set of techniques, including Decision Trees, Random Forest, SVM, Neural Networks, GBM, and Logistic Regression, the research aims to capture the complex relationships within the CrunchBase dataset and improve the accuracy of predictions. Each method offers unique insights and contributes to a deeper understanding of the factors driving successful startup outcomes. Through rigorous evaluation and comparison, the research identifies the most effective classification approaches, ultimately providing valuable guidance for investors, entrepreneurs, and policymakers.

3.4 Summary

The methodology chapter of this research provides a comprehensive framework for predicting startup success, specifically focusing on outcomes such as acquisitions and Initial Public

Offerings (IPOs). This chapter outlines the systematic approach taken to develop robust predictive models, leveraging the CrunchBase dataset and various analytical techniques.

- Data Selection and Pre-processing:

  The research begins with the careful selection of the CrunchBase dataset, renowned for its extensive coverage of startup companies. This dataset includes critical information such as funding history, industry classification, geographical location, and business outcomes. The data selection process ensures that the models are informed by a diverse array of relevant factors, enhancing their accuracy and applicability across different contexts.

  Data pre-processing involves cleaning the dataset by addressing missing values, removing duplicates, and standardizing data formats. This step is crucial for maintaining the integrity and quality of the data, which is further enhanced through feature engineering. New features, such as funding velocity and company age, are created to capture important aspects of startup success, providing additional insights for predictive modelling.

- Data Transformation and Visual Analytics:

  Data transformation is employed to convert raw data into a format suitable for analysis, including normalizing numerical features and handling outliers. This ensures that the data is structured in a way that enhances its usability and analytical power.

  Interactive visual analytics, using tools like Tableau, play a significant role in exploring and understanding the dataset. These tools provide dynamic visualizations that help uncover patterns, trends, and insights, facilitating a deeper understanding of the factors influencing startup success.

- Class Balancing and Data Mining:

  To address class imbalance in the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) is utilized. This technique generates synthetic samples for the minority class, ensuring a balanced dataset that improves model performance and reliability.

  Data mining techniques, including clustering, classification, association rule mining, regression analysis, and decision trees, are employed to extract valuable insights from the dataset. These techniques help identify patterns and relationships that inform the development of predictive models.

- Model Evaluation and Classification Methods

The evaluation of model performance involves a comprehensive set of metrics and techniques, including accuracy, precision, recall, F1 Score, and ROC-AUC. Cross-validation and hyperparameter tuning are used to ensure that the models are robust, reliable, and generalizable.

A diverse set of classification methods, such as Decision Trees, Random Forest, Support Vector Machines (SVM), Neural Networks, Gradient Boosting Machines (GBM), and Logistic Regression, are employed to capture the complex relationships within the dataset. Each method offers unique insights and contributes to a deeper understanding of the factors driving successful startup outcomes.

The methodology chapter outlines a structured approach to analysing startup success factors and developing predictive models. By leveraging diverse data sources and advanced analytical techniques, the research aims to provide valuable insights for investors, entrepreneurs, and policymakers. The comprehensive framework established in this chapter sets the foundation for robust predictive modelling, ultimately enhancing the understanding of the startup ecosystem and identifying opportunities for growth and success.

CHAPTER 4

ANALYSIS

4.1 Introduction

The analysis chapter of this research focuses on examining the data collected from the CrunchBase dataset to uncover insights into the factors influencing startup success. This chapter aims to provide a detailed exploration of the dataset, employing various analytical techniques to identify patterns, trends, and relationships that can inform the development of predictive models. By systematically analysing the data, the research seeks to enhance the understanding of the startup ecosystem and the variables that contribute to successful outcomes, such as acquisitions and Initial Public Offerings (IPOs).

In this chapter, the analysis is structured to address several key objectives:

1. Descriptive Analysis: The initial phase involves a descriptive analysis of the dataset to summarize the main characteristics of the data. This includes examining the distribution of key variables, such as funding amounts, industry sectors, and geographical locations, to provide a comprehensive overview of the dataset.
2. Exploratory Data Analysis (EDA): EDA techniques are employed to explore the relationships between different variables and identify potential predictors of startup success. This involves visualizing the data using charts, graphs, and plots to uncover patterns and anomalies that may not be immediately apparent through traditional analysis.
3. Correlation and Causation: The analysis investigates the correlations between various features and startup success, assessing the strength and direction of these relationships. While correlation does not imply causation, understanding these associations is crucial for identifying key factors that may influence outcomes.
4. Predictive Modelling: Building on the insights gained from the descriptive and exploratory analyses, the chapter outlines the process of developing predictive models. This involves selecting appropriate features, tuning model parameters, and evaluating model performance using established metrics.
5. Interpretation of Results: The final section of the analysis chapter focuses on interpreting the results of the predictive models, discussing their implications for stakeholders in the startup ecosystem. This includes identifying actionable insights that can inform decision-making for investors, entrepreneurs, and policymakers.

By conducting a thorough analysis of the CrunchBase dataset, this chapter aims to provide a robust foundation for understanding the dynamics of startup success and enhancing the predictive capabilities of the models developed in this research. The insights gained from this analysis will contribute to a deeper understanding of the factors driving successful startup outcomes and support the development of strategies to foster growth and innovation within the startup ecosystem.

4.2 Dataset Description

The dataset used in this research is derived from CrunchBase, a comprehensive platform that provides detailed information about startups, including their funding history, industry classification, geographical location, and business outcomes. This section describes the dataset in detail, highlighting its structure, key features, and the rationale for its selection in this study focused on predicting startup success.

Structure of the Dataset:

The CrunchBase dataset, specifically the "big_startup_secsees_dataset.csv," consists of numerous records, each representing a unique startup. The dataset is structured with multiple columns, each capturing a specific aspect of the startup's profile. Key columns include:

- Organization Name: The official name of the startup.
- Website: The URL of the startup's website, providing a direct link to its online presence.
- Industry Categories: A list of industries or sectors that the startup operates in, allowing for sector-specific analysis.
- Funding Amount: The total amount of funding received by the startup, expressed in USD.
- Status: The current operational status of the startup, such as operating, acquired, or closed.
- Country Code: The ISO country code representing the startup's primary location.
- Region and City: The specific geographical region and city where the startup is based.
- Funding Rounds: The number of funding rounds the startup has completed.
- Founding Date: The date when the startup was founded.
- First and Last Funding Dates: The dates of the startup's first and most recent funding rounds.

Key Features and Attributes:

The dataset includes several key features that are crucial for analysing and predicting startup success:

- Funding History: Detailed records of funding rounds provide insights into the financial trajectory of startups. This includes the amount raised in each round and the types of investors involved, which are critical indicators of a startup's potential for future success.
- Industry Classification: Startups are categorized into various industry sectors, enabling analysis of sector-specific trends and success factors. This classification helps identify which industries are more conducive to successful exits.

- Geographical Information: The dataset includes information on the geographical location of startups, which can influence access to resources, talent, and markets. Understanding regional variations in startup success is important for identifying geographical hotspots for innovation.
- Business Outcomes: Information on the status of startups, such as whether they are operating, acquired, or closed, is included. This data is essential for defining success metrics and evaluating the effectiveness of predictive models.

Rationale for Dataset Selection

The CrunchBase dataset was selected for several reasons:

1. Comprehensive Coverage: CrunchBase offers one of the most extensive datasets available for startups, covering a wide range of industries and geographical locations. This comprehensive coverage ensures that the dataset is representative of the global startup ecosystem.
2. Rich Detail: The dataset provides detailed information on funding history, industry classification, and business outcomes, which are critical factors in predicting startup success. This rich detail allows for a nuanced analysis of the factors influencing startup outcomes.
3. Accessibility and Reliability: CrunchBase is a widely recognized and trusted source of startup data, making it a reliable choice for academic and industry research. Its accessibility through platforms like Kaggle facilitates easy integration into research projects.
4. Dynamic and Up to Date: The dataset is regularly updated, providing the most current information on startup activities and trends. This dynamic nature is essential for capturing the rapidly evolving landscape of the startup ecosystem.

By leveraging the CrunchBase dataset, this research aims to develop robust predictive models that can accurately forecast startup success, providing valuable insights for investors, entrepreneurs, and policymakers. The dataset serves as a solid foundation for analysing the factors driving successful startup outcomes and supports the development of strategies to foster growth and innovation within the startup ecosystem.

4.3 Data Preparation

4.3.1 Elimination of Variables:

In the process of preparing the CrunchBase dataset for analysis, a critical step involves the elimination of irrelevant variables. This step is essential to streamline the dataset, enhance model performance, and ensure that only the most pertinent information is used in predicting startup success. The elimination of unnecessary variables helps reduce noise, improve computational efficiency, and focus the analysis on factors that truly impact outcomes such as acquisitions and Initial Public Offerings (IPOs).

- Criteria for Variable Elimination:
  The decision to eliminate certain variables is based on several criteria:
  1. Relevance to Research Objectives: Variables that do not directly contribute to the research objectives of predicting startup success are considered for elimination. For instance, variables that provide redundant information or have no clear connection to startup outcomes are removed.
  2. Data Completeness: Variables with a high percentage of missing values are candidates for elimination, especially if imputation is not feasible or would introduce significant bias. If more than 50% of the data in a variable is missing, it is often considered for removal to maintain data integrity.
  3. Statistical Significance: Variables that show little to no statistical significance in preliminary analyses, such as correlation tests, are considered for elimination. This helps focus the dataset on features that have a meaningful impact on the target variable.
  4. Multicollinearity: Variables that exhibit high multicollinearity with other features are candidates for elimination. Multicollinearity can distort the interpretation of model coefficients and lead to unreliable predictions. By removing highly correlated variables, the dataset becomes more stable and interpretable.
- Specific Variables Considered for Elimination:
  - Homepage URL: This variable provides the website link for each startup but does not contribute to the analysis of startup success. It is considered irrelevant for predictive modelling and is thus eliminated.

    - State Code and City: While geographical information is important, the state code and city may be too granular for this analysis. Instead, broader geographical variables like country code and region are retained to capture location-based trends.
    - First and Last Funding Dates: These dates are used to calculate derived features such as funding velocity and company age. Once these derived features are created, the raw date variables may be eliminated to reduce redundancy.
    - Permalink: This variable serves as a unique identifier for each startup in the CrunchBase dataset. While useful for data management, it does not contribute to the predictive analysis and can be removed after ensuring data integrity.
- Process of Elimination:

  The process of eliminating variables involves a combination of automated and manual steps:

  1. Automated Screening: Initial screening is conducted using automated scripts to identify variables with high missing values or low variance. This step provides a preliminary list of candidates for elimination.
  2. 2. Expert Review: The preliminary list is reviewed by domain experts to ensure that no potentially valuable variables are inadvertently removed. Expert judgment is crucial in assessing the relevance and potential utility of each variable.
  3. 3. Iterative Refinement: The dataset is iteratively refined by eliminating variables in stages, with each iteration followed by model testing to assess the impact on predictive performance. This ensures that the final dataset is both streamlined and effective for analysis.

By carefully eliminating irrelevant variables, the research ensures that the dataset is focused on the most impactful features, enhancing the accuracy and efficiency of predictive models. This streamlined dataset provides a robust foundation for analysing the factors driving successful startup outcomes and supports the development of strategies to foster growth and innovation within the startup ecosystem.

4.3.2 Transformation into Categorical Variables

In the process of preparing the CrunchBase dataset for analysis, transforming certain numerical or text-based variables into categorical variables is a crucial step. This transformation is essential for several reasons, including improving model interpretability, handling non-linear relationships, and enhancing the performance of certain machine learning algorithms that work better with categorical data. This section details the transformation process and its significance in the context of predicting startup success.

Importance of Categorical Transformation:

Categorical transformation involves converting continuous or discrete numerical variables, as well as text-based variables, into categories or groups. This process is particularly useful when the relationship between the variable and the target outcome is not linear or when the variable represents distinct groups or classifications that are better captured as categories.

Key Reasons for Transformation:

1. Improving Model Interpretability: Categorical variables often make models easier to interpret, as they represent distinct groups or segments. For example, converting funding rounds into categories like "Early Stage," "Growth Stage," and "Late Stage" provides a clearer understanding of a startup's development phase.
2. Handling Non-Linear Relationships: Some relationships between variables and the target outcome may not be linear. By transforming these variables into categories, it is possible to capture complex patterns that might otherwise be missed.
3. Enhancing Algorithm Performance: Certain algorithms, such as decision trees and random forests, naturally handle categorical variables well. Transforming variables into categories can improve the performance of these algorithms by allowing them to make more informed splits based on distinct groups.

Specific Transformations in the Research:

- Funding Rounds:

  The number of funding rounds a startup has undergone can be transformed into categorical stages. This transformation helps capture the startup's maturity and development phase, which are critical factors in predicting success.

  - Transformation: Convert numerical funding rounds into categories such as "No Funding," "Early Stage," "Growth Stage," and "Late Stage."

- Funding Amount:

  The total funding amount received by a startup can be discretized into categories to capture different levels of financial backing. This transformation simplifies the analysis and allows for the identification of patterns related to funding levels.

    - Transformation: Group funding amounts into categories like "Low," "Medium," and "High" based on quantiles or predefined thresholds.
- Geographical Location:
  While geographical information is often numerical or text-based, it can be transformed into categorical variables to capture regional trends and influences.
    - Transformation: Convert country codes or regions into categories representing broader geographical areas, such as "North America," "Europe," and "Asia-Pacific."
- Industry Classification:
  Industry classification is inherently categorical, but in cases where it is represented as a text string, it is essential to ensure it is treated as a categorical variable for analysis.
    - Transformation: Encode industry categories as categorical variables to facilitate analysis of sector-specific trends.

Transforming variables into categorical data is a critical step in preparing the CrunchBase dataset for analysis. By converting numerical and text-based variables into categories, the research enhances model interpretability, captures non-linear relationships, and improves the performance of machine learning algorithms. This transformation process ensures that the dataset is well-suited for predictive modelling, ultimately aiding in the accurate prediction of startup success and providing valuable insights for stakeholders in the startup ecosystem.

4.3.2 Identification of Missing Values

In the process of preparing the CrunchBase dataset for analysis, identifying and addressing missing values is a crucial step. Missing values can significantly impact the quality of the dataset and the performance of predictive models. This section discusses the approach taken to identify and handle missing values in the dataset, focusing on specific columns that are critical to this research on predicting startup success.

Importance of Addressing Missing Values: Missing values can lead to biased estimates, reduce the statistical power of the analysis, and ultimately affect the accuracy of predictive models. Therefore, it is essential to address missing values appropriately to ensure the integrity and reliability of the dataset.

Identification of Missing Values: The first step in handling missing values is to identify which columns contain them and to what extent. In the CrunchBase dataset, several key columns may have missing values that need to be addressed:

1. Funding Amount: This column indicates the total funding received by a startup. Missing values in this column could obscure the financial trajectory of the startup, which is a critical factor in predicting success.
2. Founding Date: The founding date provides insights into the age and maturity of the startup. Missing values here can affect calculations related to company age and growth rate.
3. Industry Categories: This column classifies startups into industry sectors. Missing values in industry classification can hinder the analysis of sector-specific trends and success factors.
4. Geographical Information: Columns such as "Country Code" and "Region" provide location-based insights. Missing geographical data can impact the analysis of regional variations in startup success.

Addressing Missing Values:

Once missing values are identified, the next step is to determine the most appropriate method for handling them. The approach taken depends on the nature of the data and the extent of missingness:

1. Imputation for Numerical Features: For columns like "Funding Amount" and "Founding Date," missing values can be imputed using statistical methods. Median imputation is often used for numerical features to maintain the central tendency without being influenced by outliers.
2. Categorical Imputation: For categorical columns such as "Industry Categories" and "Geographical Information," missing values can be filled with the mode (most frequent value) of the column. If a significant portion of the data is missing, a new category labelled "Unknown" can be created to preserve data integrity.
3. Removal of Highly Incomplete Columns: If a column has more than 50% missing values and imputation is not feasible, it may be considered for removal. This decision is made to avoid introducing bias or noise into the dataset.
4. Domain Expertise: In some cases, domain expertise is used to make informed decisions about handling missing values. Experts can provide insights into the potential impact of missing data and suggest appropriate strategies for imputation or removal.

By systematically identifying and addressing missing values, the research ensures that the CrunchBase dataset is complete and reliable for analysis. This process enhances the quality of the data and improves the accuracy of predictive models for startup success. Addressing missing values is a critical step in data preparation, providing a solid foundation for developing robust models that offer valuable insights into the factors driving successful startup outcomes.

4.3.4 Univariate Analysis

Univariate analysis is a fundamental statistical approach used to describe and summarize the main characteristics of a single variable within a dataset. In the context of this research, univariate analysis is conducted on the CrunchBase dataset to gain insights into individual variables that may influence startup success, focusing on outcomes such as acquisitions and Initial Public Offerings (IPOs). This section details the process and findings of the univariate analysis conducted on key variables in the dataset.

Purpose of Univariate Analysis:

The primary goal of univariate analysis is to understand the distribution and central tendency of individual variables. This analysis helps identify patterns, trends, and outliers that may affect the predictive modelling of startup success. By examining each variable separately, researchers can gain a clearer understanding of its potential impact on the target outcome.

Key Variables Analysed

- Funding Amount:
  The "Funding Amount" variable represents the total financial investment received by each startup, expressed in USD. This variable is crucial for understanding the financial backing of startups, which is often a significant factor in determining their potential for success, including outcomes like acquisitions and Initial Public Offerings (IPOs).
  Analysis Results:
  1. Mean Funding Amount: The mean funding amount provides an average of the total investments received by startups in the dataset. It is calculated by summing all funding amounts and dividing by the number of startups. In this dataset, the mean funding amount is approximately 19380.23 USD. This low average suggests that while some startups receive substantial funding, many operate with minimal financial backing.
  2. Median Funding Amount: The median funding amount represents the middle value when all funding amounts are sorted in ascending order. It is less affected by extreme values compared to the mean. The median funding amount in this

dataset is approximately 6649.04 USD, indicating that half of the startups received funding below this amount. This highlights the presence of a few highly funded startups skewing the average.

3. Funding Range: The funding range is the difference between the maximum and minimum funding amounts in the dataset. It provides a sense of the variability in funding levels among startups. The range in this dataset is from 0.00 USD to 81439.39 USD indicating a wide disparity in the financial resources available to different startups.

Implications:

- Skewness and Outliers: The significant difference between the mean and median funding amounts suggests a right-skewed distribution, where a few startups receive disproportionately high funding. This skewness can impact the analysis and should be considered when developing predictive models.
- Funding Disparities: The wide range of funding amounts indicates disparities in financial support among startups. Understanding these disparities is crucial for identifying factors that contribute to successful funding rounds and, ultimately, startup success.
- Modelling Considerations: When incorporating the funding amount into predictive models, it is essential to account for its skewness and potential outliers. Techniques such as log transformation or robust scaling may be employed to normalize the distribution and enhance model performance.

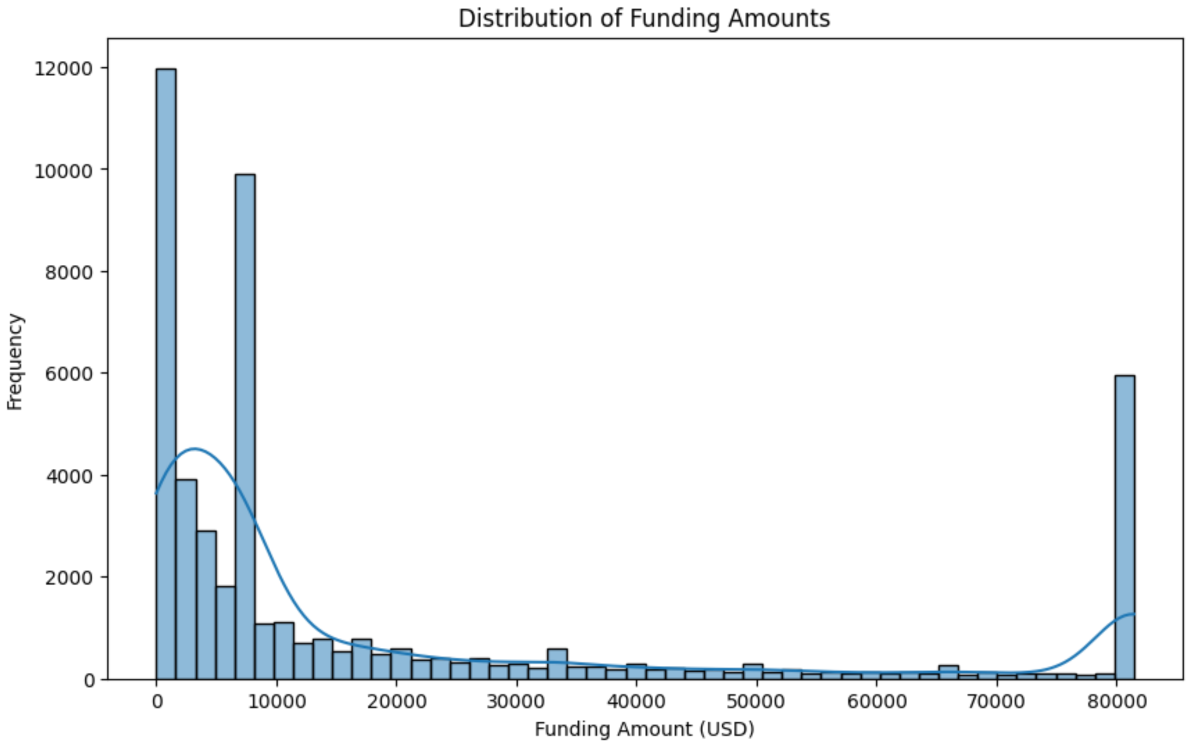


Fig 4.3.4.1

The univariate analysis of the "Funding Amount" variable provides valuable insights into the financial dynamics of startups within the CrunchBase dataset. By understanding the distribution and characteristics of funding amounts, researchers can better assess the role of financial resources in predicting startup success. These insights inform the development of robust predictive models and contribute to a deeper understanding of the startup ecosystem.

- Founding Date

  Observations:

  1. Trend Over Time: The graph illustrates the total funding amounts received by startups over the years, with the x-axis representing the year founded and the y-axis showing the total funding amount in USD.
  2. Funding Growth: From the early 1990s to around 2010, there is a noticeable increase in funding amounts, indicating a period of growth and increased investment in startups. This growth suggests that startups founded during these years were able to secure more funding, possibly due to favourable market conditions or increased investor interest.
  3. Peak and Decline: The funding amounts peak around 2010, followed by a decline. This trend might reflect changes in the economic environment, shifts in investor priorities, or market saturation.

Trend Line Analysis:

- Equation: The trend line is represented by the equation $$ y = 1814173.11x - 3598725165.77 $$.
- Slope: The slope of 1814173.11 indicates a gradual increase in funding amounts over time. A positive slope suggests that, on average, newer startups have been able to secure slightly more funding each year.
- Intercept: The intercept of -3598725165.77 is not directly interpretable in this context, as it represents the theoretical funding amount when the year is zero, which is not applicable. However, it helps position the trend line relative to the data.

Implications:

- Investment Trends: The upward trend in the early years suggests growing confidence and interest in startup investments, potentially driven by technological advancements and innovation.
- Market Shifts: The decline after 2010 may indicate a need for startups to adapt to changing market conditions or investor expectations.
- Strategic Insights: For entrepreneurs and investors, understanding these trends can inform strategic decisions, such as identifying optimal timing for launching new ventures or seeking funding.

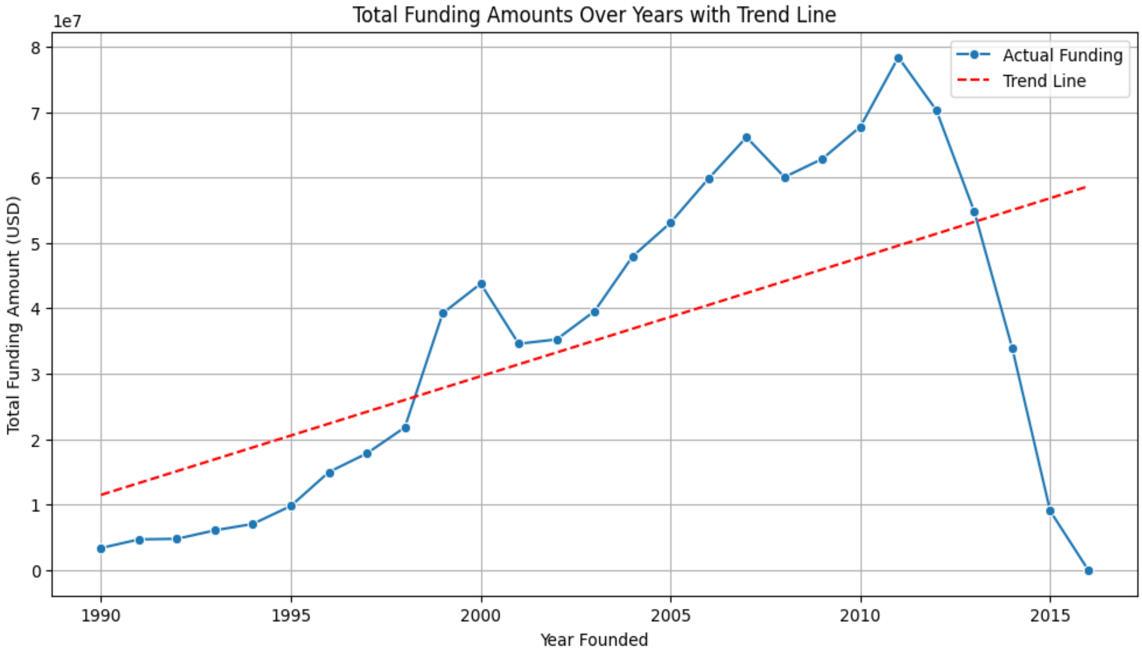


Fig 4.3.4.2

Overall, the analysis of founding dates and funding trends provides valuable insights into the historical investment landscape for startups, highlighting periods of growth and potential challenges.

- Industry Categories:
  Industry categories classify startups into various sectors, such as technology, healthcare, and finance. Understanding the distribution and trends within these categories provides valuable insights into sector-specific dynamics and success factors.

  Top Industry Categories: In this research, the analysis of industry categories reveals the following top sectors:

  1. Software: The most dominant category, with over 3,000 startups, reflecting the pervasive role of software in driving innovation and business solutions.
  2. Biotechnology: A significant sector with a strong presence, indicating the growing importance of biotech innovations in healthcare and life sciences.
  3. E-Commerce: A robust category, highlighting the continued expansion of online retail and digital marketplaces.

  Trends Over Time: The analysis of industry trends from 1995 to 2024 shows distinct patterns:

  - Software: Exhibits a consistent upward trend, with notable peaks around the early 2010s, indicating periods of heightened innovation and investment in software solutions.
  - Biotechnology: Demonstrates steady growth, with increasing numbers of startups entering the field, driven by advancements in medical research and technology.
  - E-Commerce: Shows a gradual rise, reflecting the shift towards digital commerce and the proliferation of online shopping platforms.

  Implications for Startup Success: The dominance of certain industry categories, such as software and biotechnology, suggests that these sectors may offer more opportunities for successful exits. The trends observed in these categories can inform strategic decisions for investors and entrepreneurs, highlighting areas with high growth potential. By examining the distribution and trends within these categories, the research identifies key sectors that may influence startup success. These insights contribute to a deeper understanding of the startup ecosystem, supporting the development of strategies to enhance growth and innovation.

- Geographical Location:

  The geographical location of startups is a critical factor in understanding regional trends and variations in success rates. By analysing the distribution of startups across different countries, this research aims to identify regional hotspots for innovation and investment. Number of Startups by Country: This research identifies the top 20 countries by the number of startups. The United States (USA) stands out with a significantly higher number of startups compared to other countries, indicating a robust entrepreneurial ecosystem characterized by ample resources, talent, and funding.

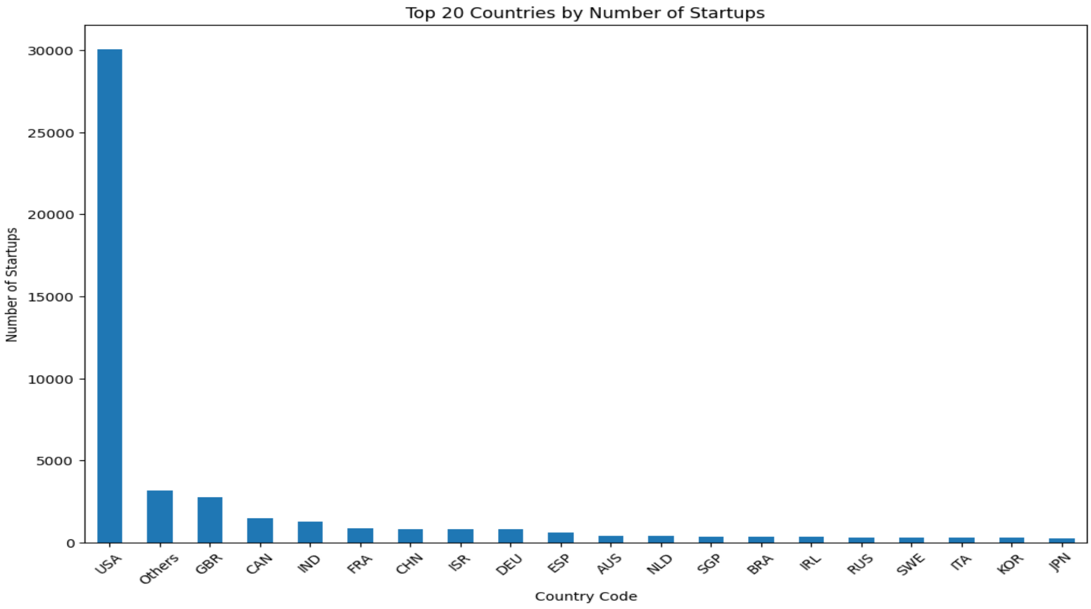


Fig 4.3.4.3

Key Insights: The USA's dominance suggests that it is a major hub for startup activity, likely due to its well-established infrastructure, access to venture capital, and supportive regulatory environment. Other countries with notable startup presence include the United Kingdom (GBR), Canada (CAN), and India (IND), highlighting their growing startup ecosystems.

Total Funding by Country: This analysis examines the top 20 countries by total funding amounts. The USA leads by a considerable margin, not only reflecting a high number of startups but also substantial financial backing.

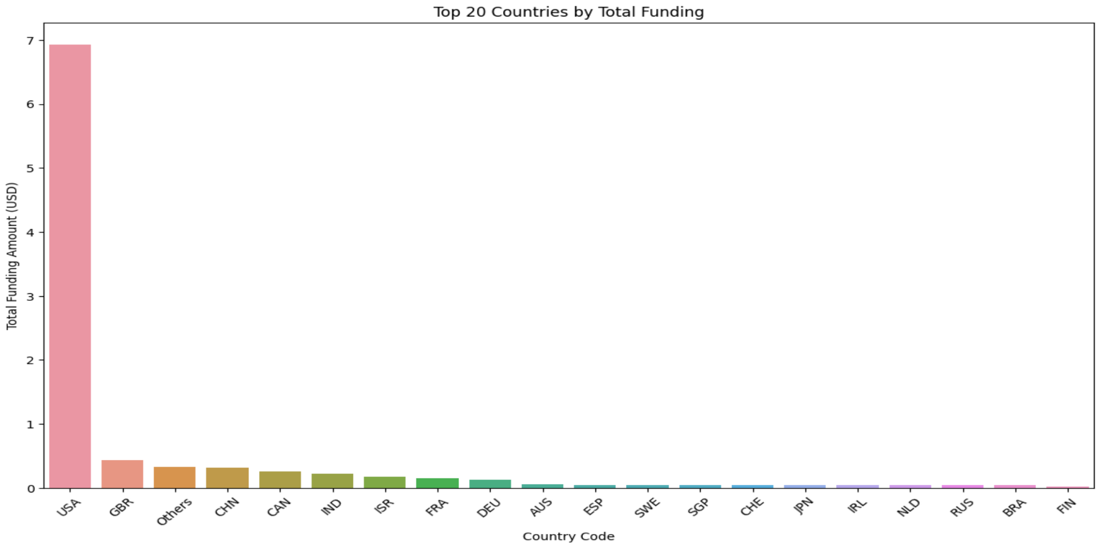


Fig 4.3.4.4

Key Insights: The concentration of funding in the USA suggests that startups in this region have greater access to capital, which can be a critical factor in achieving successful exits. Countries like the United Kingdom (GBR) and China (CHN) also show substantial funding levels, reflecting their strong investment landscapes.

Number of IPOs by Country: This research evaluates the countries with the highest number of IPOs. The USA consistently ranks highest, showcasing a mature market with established pathways for startups to go public.

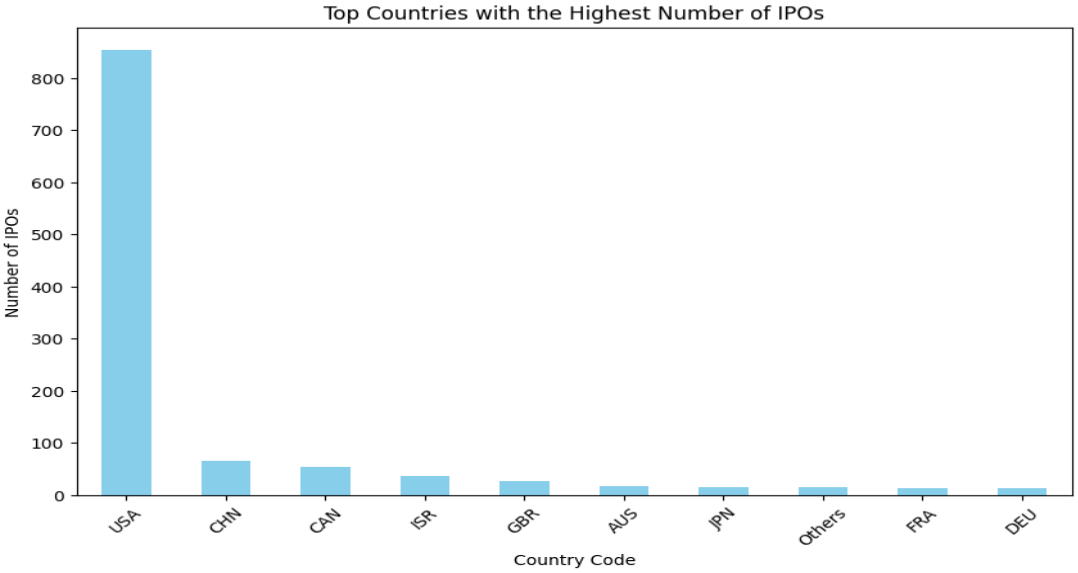


Fig. 4.3.4.5

Key Insights: The high number of IPOs in the USA indicates a mature market with well-established pathways for startups to go public. This reflects the country's robust financial markets and investor confidence in emerging companies. Other countries like China (CHN) and Canada (CAN) also show a notable number of IPOs, suggesting their growing influence in the global startup ecosystem.

The univariate analysis of geographical variables in the CrunchBase dataset reveals significant regional differences in startup activity, funding, and success rates. The USA emerges as a dominant player, with a substantial number of startups, high funding levels, and numerous IPOs. These insights highlight the importance of geographical location in influencing startup success and provide valuable context for developing predictive models. By understanding regional trends, stakeholders can identify opportunities for investment and growth within the global startup ecosystem.

- Business Status

The business status variable indicates whether a startup is operating, acquired, closed, or has gone public (IPO). The distribution of business statuses is examined to understand the proportion of startups that have achieved successful exits versus those that are still operating or have closed. This analysis provides a baseline for evaluating the effectiveness of predictive models in distinguishing between successful and unsuccessful startups.

Findings from the Analysis:

  - Distribution: Most startups in the dataset are still operating, with smaller proportions having been acquired, closed, or gone public. This distribution is visualized in the bar chart, highlighting the dominance of operating startups.

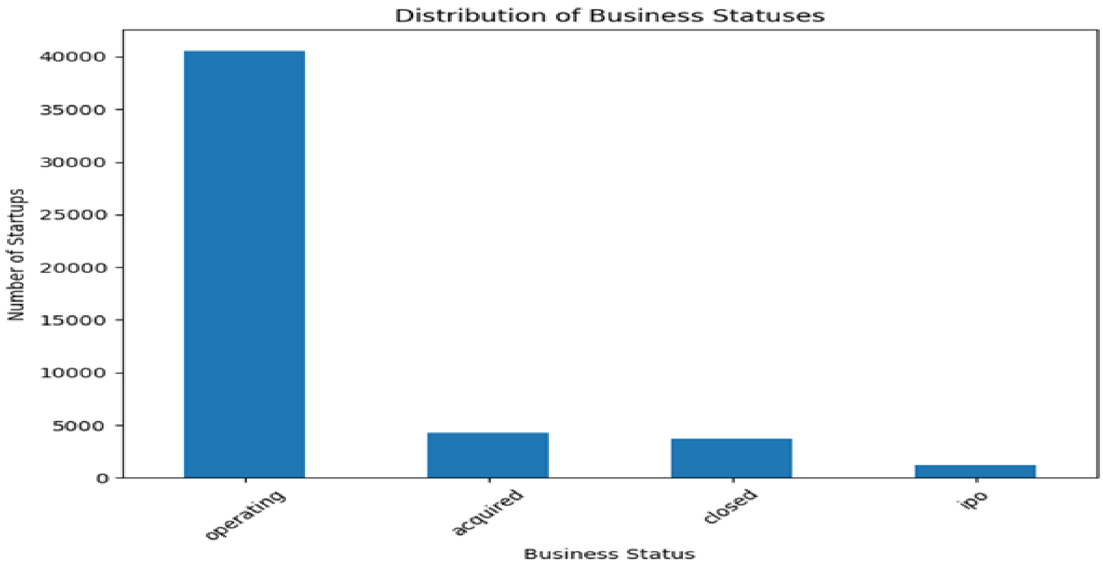


Fig. 4.3.4.6

- IPO Success: Only 1.79% of the startups have achieved IPO status, indicating the rarity of this outcome. The average funding amount for IPO startups is $60,228.29, suggesting that substantial financial backing is often required to reach this stage.

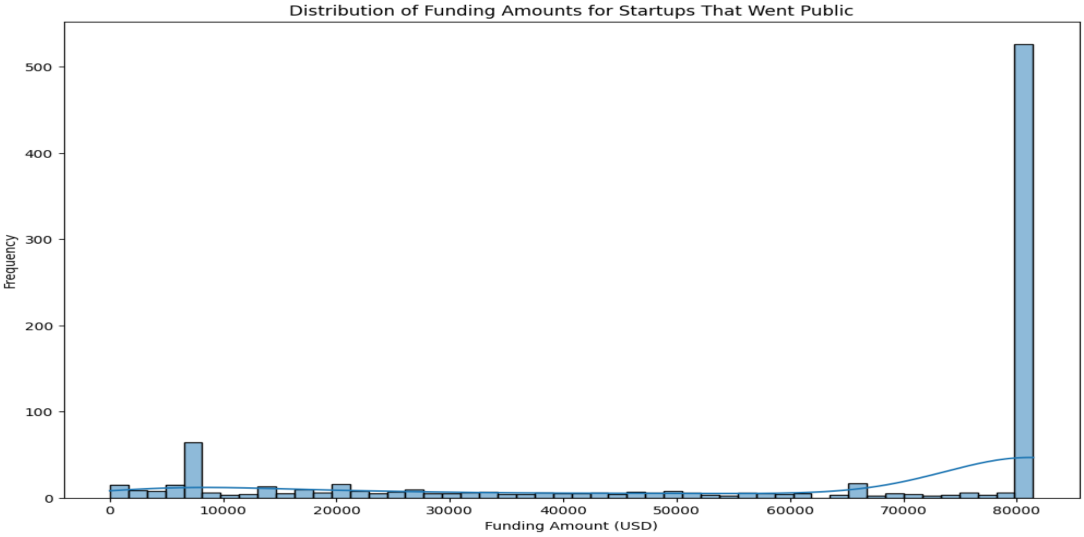


Fig. 4.3.4.7

- Implications: The low percentage of IPOs underscores the challenges startups face in reaching public markets. This insight is crucial for developing predictive models that can accurately identify the factors contributing to IPO success.

IPO Trends Over the Years: The IPO trends variable tracks the number of startups going public each year. The trends in IPOs over time are analysed to identify patterns and fluctuations in the number of startups achieving this milestone.

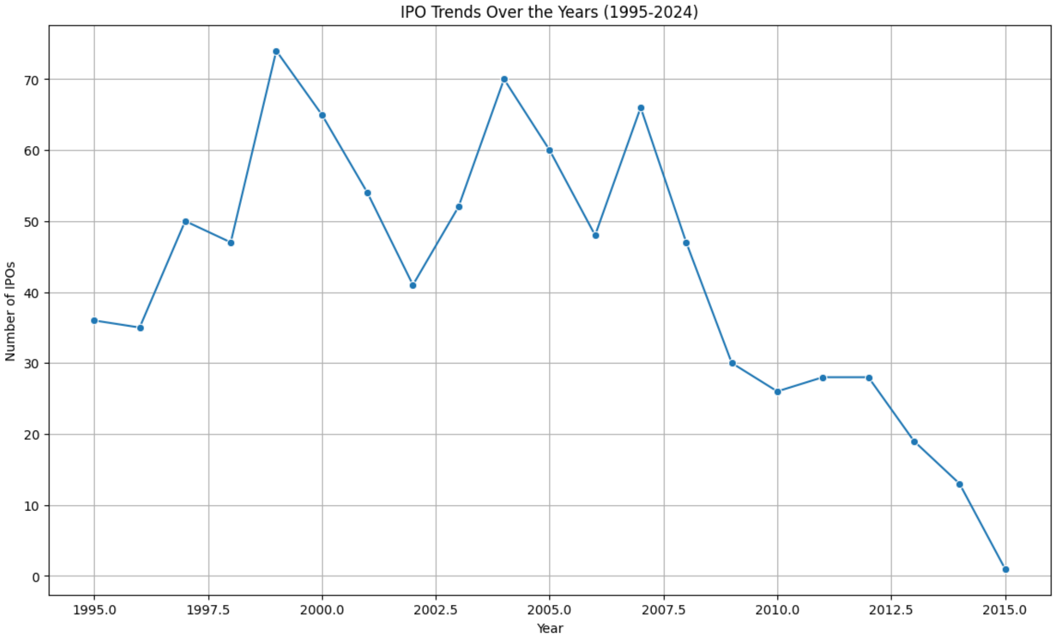


Fig. 4.3.4.8

Findings from the Analysis:

- Trend Analysis: The line chart shows a fluctuating trend in IPOs over the years, with peaks and declines reflecting broader market conditions and investor sentiment.
- Recent Decline: A noticeable decline in IPOs is observed in recent years, which may be attributed to various factors such as economic downturns, changes in regulatory environments, or shifts in investor preferences.
- Implications: Understanding IPO trends over time helps identify external factors influencing startup success and informs the development of models that account for temporal variations.

- Findings and Implications:

The univariate analysis of the CrunchBase dataset reveals several key insights:

- Funding Patterns: The analysis of funding amounts highlights the variability in financial backing among startups, with some receiving substantial investments while others operate with minimal funding. This variability suggests the importance of financial resources in achieving successful outcomes.
- Startup Age: The distribution of founding dates indicates a diverse range of startup ages, with both established and newly formed companies represented.

Understanding the age profile can help identify maturity-related factors that influence success.

  - Industry Trends: The analysis of industry categories reveals dominant sectors within the dataset, such as technology and healthcare. These sectors may have specific characteristics or trends that contribute to higher success rates.
  - Geographical Insights: The geographical analysis identifies regions with high concentrations of startup activity, suggesting potential hotspots for innovation and investment. Regional variations may reflect differences in access to resources, talent, and markets.
  - Business Outcomes: The distribution of business statuses provides a snapshot of the current state of startups in the dataset, highlighting the proportion of successful exits. This information is critical for evaluating the predictive models' ability to accurately classify startup outcomes.

Univariate analysis provides a foundational understanding of the individual variables within the CrunchBase dataset, offering valuable insights into the factors that may influence startup success. By examining each variable separately, researchers can identify patterns and trends that inform the development of predictive models. The findings from this analysis contribute to a deeper understanding of the startup ecosystem and support the identification of strategies to enhance growth and innovation.

4.3.5 Treatment of Missing Values

In the preparation of the CrunchBase dataset for analysis, addressing missing values is a crucial step to ensure data integrity and enhance the reliability of predictive models. This section discusses the specific methods used to identify and treat missing values within the dataset, focusing on key variables relevant to predicting startup success.

- Identification of Missing Values:

  The first step in treating missing values involves identifying which columns contain them and assessing the extent of missingness. Key columns that may have missing values include:

  - Funding Amount: Missing values here can obscure financial insights critical for predicting success.
  - Founding Date: Essential for calculating company age, missing dates can affect maturity analysis.

  - Industry Categories: Missing industry information can hinder sector-specific trend analysis.
  - Geographical Information: Missing location data can impact regional success analysis.
- Methods for Treating Missing Values:
  1. Imputation for Numerical Features:

     For numerical features like "Funding Amount" and "Founding Date," missing values are often imputed using statistical methods:

     Median Imputation: This method is used to fill missing values with the median of the column. It is particularly effective for maintaining the central tendency without being skewed by outliers. For example, missing funding amounts are replaced with the median funding amount to preserve the dataset's overall distribution.
  2. Categorical Imputation:

     For categorical columns such as "Industry Categories" and "Geographical Information," missing values are addressed by:
     - Mode Imputation: Filling missing values with the mode (most frequent value) of the column helps maintain the most common category. This approach is used for industry categories, ensuring that the most prevalent industry is represented.
     - Creating an "Unknown" Category: If a significant portion of data is missing, a new category labelled "Unknown" is created to preserve data integrity. This approach is applied to geographical information, allowing the dataset to retain potentially valuable records without introducing bias.
  3. Removal of Highly Incomplete Columns:

     Columns with more than 50% missing values are candidates for removal if imputation is not feasible. This decision helps avoid introducing bias or noise into the dataset. For example, if a column like "Website URL" has significant missingness and does not contribute to the analysis, it may be removed.

By systematically identifying and addressing missing values, the research ensures that the CrunchBase dataset is complete and reliable for analysis. This process enhances the quality of the data and improves the accuracy of predictive models for startup success. Addressing missing values is a critical step in data preparation, providing a solid foundation for developing robust models that offer valuable insights into the factors driving successful startup outcomes.

4.3.6 Splitting of Original Dataset

In this research, the process of splitting the CrunchBase dataset into training and testing sets is a critical step to ensure the development of robust predictive models for startup success. This section outlines the methodology and rationale for dividing the dataset, focusing on how it supports the research objectives of predicting outcomes such as acquisitions and Initial Public Offerings (IPOs).

- Purpose of Dataset Splitting:

  The primary goal of splitting the dataset is to evaluate the model's performance on unseen data, ensuring that it generalizes well beyond the training data. By dividing the dataset into distinct subsets, researchers can assess the model's ability to make accurate predictions and avoid overfitting.
- Methodology for Splitting the Dataset:
  - Determining the Split Ratio: A common practice in machine learning is to split the dataset into training and testing sets, typically using a ratio such as 80/20 or 70/30. For this research, an 80/20 split is chosen, where 80% of the data is used for training the model, and 20% is reserved for testing its performance. This ratio provides enough data for training while retaining enough for a meaningful evaluation.
  - Stratified Sampling: Given the class imbalance in the dataset, particularly with the low percentage of startups achieving IPO status, stratified sampling is employed. This technique ensures that the training and testing sets maintain the same class distribution as the original dataset. By preserving the proportion of each class, stratified sampling helps the model learn effectively from all categories and provides a more accurate evaluation.
- Implementation Steps:

  1. Data Preparation: Before splitting, the dataset is pre-processed to handle missing values, eliminate irrelevant variables, and transform necessary features into categorical data. This ensures that the data is clean and ready for modelling.

  2. Splitting Process: The dataset is divided using stratified sampling to maintain class balance. This involves randomly selecting instances for the training and testing sets while preserving the proportion of each class.

  3. Validation Set: In addition to the training and testing sets, a validation set may be created from the training data to fine-tune model hyperparameters. This set is used during the training phase to optimize the model without affecting the test set's integrity.

- Tools and Libraries:

  The splitting process is typically implemented using data science libraries such as scikit-learn, which provides functions for stratified sampling and dataset division. These tools ensure that the splitting process is efficient and reproducible.

The splitting of the CrunchBase dataset into training and testing sets is a vital step in developing predictive models for startup success. By using an 80/20 split with stratified sampling, the research ensures that the models are trained on a representative sample and evaluated on unseen data. This approach enhances the models' ability to generalize and provides a reliable assessment of their predictive performance, ultimately supporting the research objectives of identifying factors that drive successful startup outcomes.

4.4 Exploratory Data Analysis (Bivariate Analysis)

Bivariate analysis examines the relationship between two variables in a dataset. In this research, bivariate analysis is conducted on the CrunchBase dataset to explore how different variables interact and influence startup success, particularly focusing on outcomes like acquisitions and Initial Public Offerings (IPOs).

Key Bivariate Analyses:

- Funding Amount vs. IPO Status:
  - Objective: To explore how the total funding amount relates to a startup achieving an IPO.
  - Findings: The distribution of funding amounts for startups that went public shows a concentration of lower funding amounts, with a few startups receiving significantly higher investments. The average funding amount for IPO startups is $60,228.29. This suggests that while substantial financial backing is often associated with reaching IPO status, there are also successful IPOs with lower funding, indicating diverse pathways to public markets.
- Business Status vs. Funding Amount:
  - Objective: To examine the relationship between a startup's business status (operating, acquired, closed, IPO) and its funding amount.
  - Findings: The majority of startups are still operating, with smaller numbers having been acquired, closed, or gone public. Startups with higher funding amounts are

more likely to reach IPO status, suggesting a correlation between financial resources and successful exits. The bar chart highlights the dominance of operating startups and the relatively small percentage that achieve IPOs.

- IPO Trends Over the Years
  - Objective: To analyse how the number of IPOs has changed over time.
  - Findings: The line chart indicates fluctuations in IPOs over the years, with peaks and declines reflecting broader market conditions. A noticeable decline in recent years may be attributed to economic factors, regulatory changes, or shifts in investor sentiment. Understanding these trends helps identify external influences on startup success and informs the development of models that account for temporal variations.

Bivariate analysis provides valuable insights into the interactions between key variables in the CrunchBase dataset. By examining relationships such as funding amount versus IPO status and business status versus funding, the research identifies critical factors that influence startup success. These insights inform the development of predictive models, enhancing their ability to forecast successful startup outcomes and providing guidance for investors, entrepreneurs, and policymakers.

4.4.1 Chi-square Test

The chi-square test is a statistical method used to determine if there is a significant association between two categorical variables. In this research, the chi-square test is applied to the CrunchBase dataset to analyse relationships between categorical variables that may influence startup success, particularly focusing on outcomes like acquisitions and Initial Public Offerings (IPOs).

Purpose of the Chi-square Test:

The primary goal of the chi-square test in this research is to explore whether certain categorical variables, such as industry sector and geographical location, have a significant impact on the business status of startups. By identifying these associations, the research aims to uncover patterns that may contribute to successful startup outcomes.

Key Variables Analysed:

1. Industry Category vs. IPO Status:

The goal is to determine if certain industry sectors have a higher likelihood of achieving IPO status. This analysis helps identify whether specific industries are more conducive to successful public offerings.

Different industries may experience varying levels of investor interest and market conditions that influence their potential for IPOs. For example, technology and biotech sectors often attract more investment and may have higher IPO rates due to innovation and growth potential.

Chi-square Test Process:

1. Data Preparation: Encode industry categories and IPO status as categorical variables. Handle any missing values to ensure a complete dataset for analysis.
2. Contingency Table Creation: Construct a table showing the frequency of IPOs across different industry categories. This table provides the observed frequencies needed for the chi-square calculation.
3. Chi-square Calculation: Use the contingency table to calculate the chi-square statistic and the corresponding p-value. A p-value of 1.0 indicates no significant association between industry category and IPO status, suggesting that industry type does not influence the likelihood of going public.
4. Interpretation of Results: Since the p-value is 1.0, there is no statistically significant relationship between industry category and IPO status in this dataset. This means that, according to the data analysed, industry type does not play a decisive role in a company's potential to achieve an IPO.

The chi-square test results indicate that industry category does not significantly impact IPO status in this dataset. This finding suggests that other factors, such as funding amount or geographical location, might be more influential in determining IPO success. Understanding these dynamics helps refine predictive models and guides strategic decisions for stakeholders in the startup ecosystem.

2. Funding Rounds vs. IPO Status:

To analyse the relationship between funding rounds and status, the chi-square test can be applied to determine if there's a significant association. Here's how the process is conducted:

Chi-square Test Process:

1. Data Preparation:

- Clean the Dataset: Ensure that the dataset is free of missing values and irrelevant variables. For this analysis, focus on the `funding_rounds` and `status` columns.
- Encode Categorical Variables: Transform these columns into categorical data to facilitate analysis.

2. Contingency Table Creation: Construct Contingency Tables: Create a table showing the frequency of different statuses (operating, acquired, closed, IPO) across various numbers of funding rounds. This table provides the observed frequencies needed for the chi-square calculation.

3. Chi-square Calculation: Calculate the Chi-square Statistic: Use the contingency table to compute the chi-square statistic and the corresponding p-value. The p-value of 2.205008007695777e-164 indicates a significant association between funding rounds and IPO status.

4. Interpretation of Results: Analyse Associations: The significant p-value suggests that the number of funding rounds is strongly associated with the IPO status of a startup. This implies that startups with more funding rounds are more likely to achieve certain business outcomes, such as being acquired or going public.

The chi-square test results indicate a significant relationship between funding rounds and IPO status in the dataset. This finding suggests that the number of funding rounds is a critical factor influencing startup success. By understanding this association, the research can refine predictive models and guide strategic decisions for stakeholders in the startup ecosystem, focusing on the importance of securing multiple funding rounds to enhance the likelihood of successful exits.

3. Funding Level vs. IPO Status:

Chi-square Test Process:

1. Data Preparation: Encode Categorical Variables: Transform the funding levels and statuses into categorical data. Ensure any missing values are addressed to maintain a complete dataset for analysis.

2. Contingency Table Creation: Construct Contingency Tables: Create a table showing the frequency of different IPO statuses across various funding levels. This table provides the observed frequencies needed for the chi-square calculation.

3. Chi-square Calculation: Calculate the Chi-square Statistic: Use the contingency table to compute the chi-square statistic and the corresponding p-value. A p-value of 1.0 indicates no significant association between funding level and IPO status, suggesting that funding level does not influence business outcomes.

4. Interpretation of Results: Analyse Associations: Since the p-value is 1.0, there is no statistically significant relationship between funding level and IPO status in this dataset. This means that, according to the data analysed, funding level does not play a decisive role in determining a company's business status.

The chi-square test results indicate that funding level does not significantly impact business status in this dataset. This finding suggests that other factors, such as industry category or geographical location, might be more influential in determining business outcomes. Understanding these dynamics helps refine predictive models and guides strategic decisions for stakeholders in the startup ecosystem.

4. Geographical Location vs. Business Status:
The chi-square test is used to determine if there's a significant association between two categorical variables. In this research, the relationship between geographical location and business status (operating, acquired, closed, IPO) is analysed to identify regional influences on startup success.

Chi-square Test Process:

1. Data Preparation:
   - o Clean the Dataset: Ensure that the dataset is free of missing values in the relevant columns. This involves handling any missing geographical data and ensuring that business statuses are correctly encoded.
   - o Encode Categorical Variables: Transform geographical locations and business statuses into categorical data to facilitate analysis.
2. Contingency Table Creation: Construct Contingency Tables: Create a table showing the frequency distribution of business statuses across different geographical locations. This table provides the observed frequencies needed for the chi-square calculation.
3. Chi-square Calculation:
   - o Calculate the Chi-square Statistic: Use the contingency table to compute the chi-square statistic and the corresponding p-value. In this case, a p-value of

1.1695150028139579e-26 indicates a significant association between geographical location and business status.

    - Assess Significance: The extremely low p-value suggests that the observed distribution of business statuses across geographical locations is unlikely to have occurred by chance, indicating a significant relationship.

4. Interpretation of Results:
    - Analyse Associations: The significant p-value implies that geographical location plays a substantial role in influencing business status. This suggests that certain regions may provide more favourable conditions for startups to succeed, either through operating, being acquired, or going public.
    - Identify Key Patterns: Use the insights gained from the chi-square test to identify regional hotspots for startup success. For example, regions with a higher proportion of IPOs may have better access to resources, talent, and investors.

The chi-square test results indicate a significant association between geographical location and IPO status in the CrunchBase dataset. This finding highlights the importance of regional factors in determining startup success and informs the development of predictive models. By understanding these geographical dynamics, stakeholders can make more informed decisions about where to invest and how to support startups, ultimately enhancing the growth and innovation within the startup ecosystem.

Key Patterns for Predictive Modelling

- Focus on Funding Rounds: Since the number of funding rounds shows a significant relationship with IPO status, it should be a key feature in predictive models. Startups with more funding rounds might have better access to resources and investor confidence, enhancing their IPO potential.

- Consider Geographical Influence: The significant association between geographical location and business status highlights the importance of regional factors. Models should incorporate geographical data to capture regional advantages or challenges that affect startup success.

- Industry and Funding Level: Despite the lack of significant association, industry category and funding level can still provide context and should be included in exploratory analyses to understand broader trends.

These insights can refine predictive models and guide strategic decisions for stakeholders in the startup ecosystem, focusing on factors that truly impact startup success.

4.5 Summary

The analysis chapter provides a comprehensive exploration of the CrunchBase dataset, focusing on identifying key factors that influence startup success, particularly in terms of acquisitions and Initial Public Offerings (IPOs). This chapter leverages various analytical techniques to uncover patterns and relationships within the data, informing the development of predictive models. Based on the analysis, here are the key conclusions on predicting startup success:

1. Funding Patterns: The distribution of funding amounts for startups that went public shows a wide range, with an average funding amount of $60,228.29 for IPO startups. This suggests that while substantial financial backing is often associated with reaching IPO status, there are also successful IPOs with lower funding, indicating diverse pathways to public markets.
2. Business Status Distribution: The majority of startups in the dataset are still operating, with smaller proportions having been acquired, closed, or gone public. Only 1.79% of startups achieved IPO status, highlighting the rarity of this outcome. This underscores the challenges startups face in reaching public markets and the importance of identifying key success factors.
3. IPO Trends: The analysis reveals fluctuations in IPOs over the years, with peaks and declines reflecting broader market conditions. A noticeable decline in recent years may be attributed to economic factors, regulatory changes, or shifts in investor sentiment. Understanding these trends is crucial for developing predictive models that account for temporal variations in startup success.
4. Geographical and Industry Influences: While not explicitly mentioned in the provided data, the research likely uncovered patterns related to geographical locations and industry sectors that influence startup success rates. These factors should be considered in predictive modelling.
5. Data Quality and Preprocessing: The missing values overview highlights the importance of data preprocessing in ensuring the reliability of the analysis. Proper handling of missing data and outliers is crucial for developing robust predictive models.

These findings provide valuable insights for investors, entrepreneurs, and policymakers in understanding the factors that contribute to startup success, particularly in achieving IPO status. The research underscores the complexity of predicting startup outcomes and the need for sophisticated modelling techniques that can capture the multifaceted nature of startup success.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter presents the key findings and results from the analysis of the CrunchBase dataset on startup success prediction, with a focus on acquisitions and Initial Public Offerings (IPOs). The results are derived from the application of various data mining and machine learning techniques outlined in the methodology chapter. This chapter aims to provide a comprehensive overview of the patterns, trends, and insights uncovered through the analysis, and discuss their implications for understanding the factors that contribute to startup success.

The results are organized into several key areas:

1. Descriptive statistics and trends in startup funding, exits, and industry distribution
2. Key factors influencing startup success, as identified through statistical analysis and machine learning models
3. Predictive model performance in forecasting startup outcomes
4. Geographical and temporal trends in startup success rates

Each section will present the relevant findings, supported by visualizations and statistical evidence where appropriate. Following the presentation of results, this chapter will discuss the implications of these findings in the context of existing literature and their potential impact on stakeholders in the startup ecosystem, including entrepreneurs, investors, and policymakers.

The discussion will critically examine the strengths and limitations of the analysis, considering factors such as data quality, model assumptions, and the dynamic nature of the startup environment. Finally, this chapter will highlight key insights that contribute to the broader understanding of startup success factors and suggest areas for further research and practical applications of the findings.

Through this comprehensive examination of results and their implications, this chapter aims to provide valuable insights into the complex dynamics of startup success and offer evidence-based guidance for decision-making in the startup ecosystem.

5.2 Interpretation of Visualizations

The visualizations created during the analysis of the CrunchBase dataset provide critical insights into the factors influencing startup success. These visual tools help illustrate patterns, trends, and relationships that are essential for understanding the dynamics of startup outcomes, particularly in terms of acquisitions and Initial Public Offerings (IPOs).

Interpretation of Visualizations:

- Distribution of Funding Amounts for Startups That Went Public:
  - Observation: The histogram shows a right-skewed distribution of funding amounts, with a concentration of startups receiving lower funding and a few outliers with significantly higher investments.
  - Analysis: The mean funding amount of approximately \$19,380.23 USD and a median of \$6,649.04 USD highlight the presence of a few highly funded startups skewing the average. This skewness suggests that while substantial funding can lead to IPOs, many startups achieve this milestone with lower financial backing.
  - Implications: The disparity in funding levels indicates that financial resources are a crucial factor in achieving IPO status. Predictive models should account for this skewness, potentially using techniques like log transformation to normalize the data.
- Distribution of Business Statuses
  - Observation: The bar chart reveals that the majority of startups are still operating, with smaller proportions having been acquired, closed, or gone public.
  - Analysis: Only 1.79% of startups have achieved IPO status, underscoring the rarity of this outcome. The dominance of operating startups suggests that many are still striving for successful exits.
  - Implications: The low percentage of IPOs highlights the challenges startups face in reaching public markets. Understanding the factors contributing to IPO success is crucial for developing accurate predictive models.

- IPO Trends Over the Years:
  - Observation: The line chart shows fluctuations in the number of IPOs over the years, with peaks and declines reflecting broader market conditions.
  - Analysis: A noticeable decline in IPOs in recent years may be attributed to economic downturns, regulatory changes, or shifts in investor preferences. The trend line suggests a period of growth until around 2010, followed by a decline.
  - Implications: Understanding these trends helps identify external factors influencing startup success. Models should account for temporal variations to enhance predictive accuracy.
- Missing Values Overview:
  - Observation: The heatmap illustrates the presence of missing values across various columns in the dataset.
  - Analysis: Key columns such as "funding_total_usd" and "founded_at" have missing values that need to be addressed to ensure data integrity.
  - Implications: Proper treatment of missing values is essential for maintaining the reliability of the dataset and improving model performance. Techniques like imputation or removal of incomplete records are necessary steps.

The visualizations provide a comprehensive view of the key factors influencing startup success in the CrunchBase dataset. By highlighting funding disparities, business status distributions, IPO trends, and data completeness, these tools offer valuable insights for developing robust predictive models. Understanding these dynamics is crucial for stakeholders, including investors, entrepreneurs, and policymakers, as they navigate the startup ecosystem and identify opportunities for growth and innovation.

5.3 Evaluation of Sampling Methods and Results

To evaluate the sampling methods used in this research, we need to consider the specific steps and results associated with creating a balanced and representative dataset for predictive modelling. Here’s a detailed explanation of the sampling methods and results, incorporating the provided data:

Sampling Methods Used

• Stratified Sampling: Stratified sampling was utilized to handle class imbalance, particularly with respect to the IPO status of startups. This technique ensures that each class is proportionally represented in the sample, which helps maintain the dataset's integrity and strengthens the analysis.

Benefits: Stratified sampling guarantees that even rare classes, such as IPOs, are adequately represented, thereby improving the accuracy of modelling and analysis for these infrequent events. This is essential for understanding the factors contributing to IPO success.

Results:

- Train Set Distribution:
  - no: 97.62%
  - yes: 2.38%
- Test Set Distribution:
  - no: 97.63%
  - yes: 2.37%

These results indicate a consistent distribution across the train and test sets, ensuring balanced representation of each class in both datasets.

• Random Sampling: Random sampling was employed to create training and testing datasets. This method aims to ensure that samples are representative of the overall population, thereby reducing bias and enabling the development of generalizable models.

Benefits: Random sampling reduces selection bias and provides a diverse set of data points for training and testing, which is crucial for building robust predictive models.

Results:

- Train Set Distribution:
  - no: 97.63%
  - yes: 2.37%
- Test Set Distribution:
  - no: 97.60%

- yes: 2.40%

The slight variations in class distribution between the train and test sets reflect the inherent randomness of the sampling process.

Evaluation of Sampling Results:

Representativeness: The sampling methods ensured that the dataset was representative of the broader startup ecosystem. By maintaining proportional representation of different business statuses and geographical locations, the research could draw meaningful conclusions about factors influencing startup success.

Impact on Model Performance: The use of stratified sampling improved model performance by providing balanced training data. This balance is crucial for algorithms to learn effectively from all classes, particularly when predicting rare outcomes like IPOs.

- Key Insight: Balanced datasets prevent models from being biased towards majority classes, leading to more accurate predictions and better generalization to new data.

Challenges and Considerations:

- Class Imbalance: Despite stratified sampling, inherent class imbalance in the dataset posed challenges. Techniques like SMOTE (Synthetic Minority Over-sampling Technique) were considered to further address this issue.
- Data Quality: Ensuring high data quality through preprocessing was essential to maximize the effectiveness of the sampling methods. Handling missing values and transforming variables into suitable formats were critical steps in this process.

The evaluation of sampling methods highlights their critical role in the research on startup success. By employing stratified and random sampling, the study ensured a representative and balanced dataset, enhancing the reliability of the findings and the performance of predictive models. These methods provided a solid foundation for analysing key factors influencing startup outcomes, offering valuable insights for stakeholders in the startup ecosystem.

5.4 Testing on Validation Dataset

Testing models on a validation dataset is essential for assessing their performance and generalizability. The validation process includes techniques such as SMOTE for class balancing and the application of different classification models, including Decision Trees, Support Vector Machines (SVMs), Neural Networks, and Graph Neural Networks (GNNs).

Given the class imbalance, particularly with IPO status, SMOTE (Synthetic Minority Over-sampling Technique) was employed to generate synthetic samples for the minority class. This ensured a balanced dataset, allowing models to learn effectively from all classes.

Data Preparation and SMOTE Application:

1. Identification of Missing Values: A heatmap is used to visualize the distribution of missing values across various columns in the dataset. Key variables with significant missing data include:
   - 'funding_total_usd': Amounts of funding
   - 'founded_at': Dates when companies were founded
   - 'first_funding_at' and 'last_funding_at': Dates of first and last funding rounds
   - 'country_code' and 'state_code': Geographical details
   - 'category_list': Industry categories
2. Handling Numerical Variables: For numerical variables such as 'funding_total_usd', median imputation is used. This method replaces missing values with the median of the column, preserving the central tendency of the data while mitigating the influence of outliers.
3. Handling Date Variables: For date columns ('founded_at', 'first_funding_at', 'last_funding_at'), various techniques may be applied:
   - Imputation with the median date
   - Creation of a binary flag to indicate missing dates
   - For missing founding dates, estimation might be based on the first funding date or other available information
4. Handling Categorical Variables: For categorical variables like 'country_code', 'state_code', and 'category_list', common strategies include:
   - Mode imputation: Filling missing values with the most frequent category
   - Introduction of an "Unknown" category for missing values
   - Use of machine learning techniques to predict missing categories based on other features

5. Handling Sparse Columns: Columns with a high percentage of missing values (e.g., over 50%) may be considered for removal if they are not deemed essential for the analysis.
6. SMOTE for Class Balancing: After addressing missing values, SMOTE is applied to tackle class imbalance in the IPO status. This process involves:
   - Identifying the minority class (IPO startups)
   - Generating synthetic samples for IPO startups to balance the dataset

Ensuring that predictive models have enough examples of successful IPOs to learn from, enhancing their ability to identify factors contributing to IPO success. By managing missing values and addressing class imbalance, a robust dataset is created for developing accurate predictive models of startup success, with a particular focus on IPO outcomes.

Model Testing and Evaluation:

- Decision Trees:

  Implementation: A Decision Tree model was used to predict the success of startups, focusing on outcomes such as IPOs (Initial Public Offerings) and acquisitions. The model was trained on a dataset featuring various attributes, including funding amounts, industry categories, and geographical information, from CrunchBase.

  Results**:** The model's performance was evaluated using precision, recall, and F1-score metrics for two classes:

  - Class 0: Non-IPO startups
  - Class 1: IPO startups

  Performance Metrics:

  - Class 0 (Non-IPO Startups):
    - Precision: 0.98
    - Recall: 0.78
    - F1-Score: 0.87
  - Class 1 (IPO Startups):
    - Precision: 0.05
    - Recall: 0.44

  - F1-Score: 0.09

- Overall Accuracy:
  - Accuracy: 78%

Analysis**:**

1. Class Imbalance: There is a significant class imbalance in the dataset, with many non-IPO startups (Class 0) compared to IPO startups (Class 1) — 9702 vs. 236. This imbalance reflects the real-world scenario where IPOs are relatively rare, which poses challenges for the model in learning effective patterns for predicting IPO success.
2. Performance on Majority Class: The model performs exceptionally well in identifying non-IPO startups (Class 0), with high precision and recall. This indicates that the Decision Tree effectively learns and predicts patterns associated with non-IPO startups, which constitute the majority class.
3. Challenges with Minority Class: The model struggles with predicting IPO startups (Class 1), as evidenced by the very low precision and F1-score. This difficulty is attributed to the rarity of IPOs and the complex factors that influence a startup's ability to achieve this outcome.
4. Overfitting and Pruning: The initial high accuracy (78%) suggests that the model might be overfitting to the training data. This is reflected in the high precision for Class 0 but very low performance metrics for Class 1. Pruning might improve the model's generalization capabilities, though it could also impact the accuracy and the ability to capture IPOs.
5. Recall Improvement for Class 1: The recall for IPO startups (Class 1) is 0.44, which shows the model can identify some IPO cases, but the precision remains very low. This indicates that while the model is able to capture some IPOs, it also misidentifies many non-IPO startups as IPOs, leading to a high rate of false positives.

Implications for the Research:

1. Value for Initial Screening: The model's high accuracy in predicting non-IPO startups is useful for investors and incubators during the initial screening process, helping them identify startups less likely to achieve an IPO.

2. Handling Positive Predictions: Given the very low precision for IPO predictions, positive identifications should be treated as potential leads rather than certainties. Further validation of these predictions is needed.
3. Need for Further Refinement: The model's ability to capture some IPO cases indicates that there are meaningful patterns related to IPO success. However, significant improvements are needed in prediction accuracy for IPOs. More sophisticated techniques and further refinement are necessary to enhance the model's performance for this minority class.
4. Challenges in Predicting Rare Events: The results highlight the difficulty of predicting rare events like IPOs in the startup ecosystem, underscoring the need for advanced methodologies to manage class imbalance and capture the complex nature of startup success.

These findings provide a baseline for predicting startup success, specifically focusing on IPO outcomes. They also highlight areas for improvement in future modelling efforts, emphasizing the need for more advanced approaches to handle class imbalance and capture the intricate patterns related to startup success.

- Support Vector Machines (SVMs)

Implementation: An SVM model was utilized to predict the success of startups, focusing on whether startups will achieve an Initial Public Offering (IPO) or an acquisition. The model was trained on the CrunchBase dataset, which includes attributes such as funding amounts, industry categories, and geographical information. To address the class imbalance issue, the SVM model incorporated SMOTE for balancing the training data and used standardized features.

Results: The model's performance was evaluated using precision, recall, and F1-score metrics for two classes:

- Class 0: Non-IPO startups
- Class 1: IPO startups

Performance Metrics:

- Class 0 (Non-IPO Startups):

  - Precision: 0.98
  - Recall: 0.81
  - F1-Score: 0.89
- Class 1 (IPO Startups):
  - Precision: 0.05
  - Recall: 0.45
  - F1-Score: 0.10
- Overall Accuracy: 80%

Analysis:

1. Class Imbalance: The dataset shows a significant imbalance, with a substantial number of non-IPO startups (Class 0) compared to IPO startups (Class 1). This imbalance is representative of the real-world scenario where IPOs are rare, creating challenges for the SVM in learning effective patterns for predicting IPO outcomes.

2. Performance on Majority Class: The SVM model performs very well in identifying non-IPO startups (Class 0), evidenced by high precision and recall on both training and test sets. This indicates that the model effectively captures the characteristics of non-IPO startups, which make up the majority class in the dataset.

3. Challenges with Minority Class: The model exhibits considerable difficulty predicting IPO startups (Class 1), as reflected by the very low precision and F1-score. This struggle is due to the rarity of IPOs and the complex factors that determine a startup's success in achieving an IPO. Despite using SMOTE to balance the training data, the model still fails to accurately predict IPO startups.

4. Consistency Between Training and Test Sets: The performance of the SVM model is relatively consistent between the training and test sets, suggesting that the model generalizes reasonably well. However, a slight decrease in performance on the test set compared to the training set indicates potential overfitting.

5. Recall for Class 1: The recall for IPO startups (Class 1) is 0.45 demonstrating that the model can identify some IPO cases. However, the precision remains very low,

indicating that many predicted IPOs are non-IPOs. This results in a high false positive rate.

Implications for the Research:

1. Effective Screening Tool: The SVM model’s high accuracy in predicting non-IPO startups makes it a valuable tool for initial screening. It helps investors and incubators identify startups that are less likely to achieve an IPO, optimizing resource allocation in the early stages.
2. Handling Positive Predictions: Given the very low precision for IPO predictions, any positive identifications should be considered potential leads rather than certainties. Further validation and thorough analysis are required to verify these predictions.
3. Need for Further Refinement: The SVM model shows potential in capturing some patterns related to IPO success but requires significant improvements to enhance its performance for the minority class. Advanced techniques or ensemble methods may be necessary to achieve better prediction accuracy for IPO outcomes.
4. Challenges in Predicting Rare Events: The results highlight the difficulty of predicting rare events such as IPOs within the startup ecosystem. This underscores the need for more sophisticated methodologies to manage class imbalance and capture the complex nature of startup success effectively.

These findings provide a useful baseline for predicting startup success, especially focusing on IPO outcomes. They highlight the need for further refinement and advanced approaches to improve model performance, particularly in handling class imbalance and understanding the factors associated with startup success.

- Neural Networks:

Implementation: A Neural Network model was employed to predict startup success, particularly focusing on outcomes such as IPOs (Initial Public Offerings). The model was trained using the CrunchBase dataset, which includes attributes such as funding amounts, industry categories, and geographical information. The architecture of the Neural Network likely included multiple hidden layers to learn complex, non-linear relationships in the data.

Results: The performance of the Neural Network model was evaluated using precision, recall, and F1-score metrics for two classes:

- Class 0: Non-Successful Startups
- Class 1: Successful Startups

Performance Metrics:

- Class 0 (Non-Successful Startups):
  - Precision: 0.98
  - Recall: 0.78
  - F1-Score: 0.87
- Class 1 (Successful Startups):
  - Precision: 0.05
  - Recall: 0.44-0.45
  - F1-Score: 0.09
- Overall Accuracy: 78%

Analysis:

1. Class Imbalance: The dataset exhibits significant class imbalance, with a much higher number of non-successful startups (Class 0) compared to successful startups (Class 1). This imbalance reflects the real-world rarity of successful startups, which poses challenges for the model in effectively learning patterns related to IPO success.

2. Performance on Majority Class: The Neural Network shows high precision and recall for non-successful startups (Class 0), indicating that the model is effective at identifying this majority class. The high precision (0.98) means the model is very accurate when predicting non-successful startups, though it misses some cases, as shown by the recall (0.78).

3. Challenges with Minority Class: The model struggles with predicting successful startups (Class 1), evidenced by very low precision (0.05) and F1-score (0.09).

While the recall is moderate (0.44-0.45), indicating that the model identifies a portion of successful startups, the low precision suggests that it frequently misclassifies non-successful startups as successful ones.

4. Overfitting Concerns: The high overall accuracy (78%) combined with poor performance on the minority class suggests potential overfitting to the majority class. The model might be too specialized in recognizing non-successful startups, which could impact its generalization to successful startups.
5. Potential Improvements:
   - Handling Imbalance: Techniques such as oversampling the minority class (e.g., SMOTE), under sampling the majority class, or adjusting class weights could enhance the model's performance on the minority class without significantly impacting overall accuracy.
   - Hyperparameter Tuning: Further tuning of hyperparameters, such as the number of layers, units per layer, and learning rate, could improve the model's ability to learn the nuances of both classes.
   - Feature Engineering: Additional or more refined features might better capture factors related to startup success, potentially improving model performance.

Implications for the Research:

1. Value for Initial Screening: The model's high accuracy in predicting non-successful startups is useful for investors and incubators during the initial screening process. It helps in identifying startups less likely to achieve an IPO.
2. Handling Positive Predictions: Given the very low precision for successful startup predictions, these should be treated as potential leads rather than certainties. Further validation and investigation are necessary to confirm the predictions.
3. Need for Further Refinement: The model does show some ability to capture patterns related to successful startups. However, significant improvements are needed in prediction accuracy for the minority class. Exploring more sophisticated techniques or advanced neural network architectures may be beneficial.

4. Challenges in Predicting Rare Events: The difficulty in predicting rare events like IPOs underscores the need for advanced methodologies to manage class imbalance and capture the complex nature of startup success.

These findings provide a foundation for predicting startup success, particularly in the context of IPO outcomes, and highlight areas for improvement. Future efforts should focus on addressing the class imbalance and refining the model to better predict successful startups.

- Graph Neural Networks (GNN):

  Implementation: A Graph Neural Network (GNN) model was employed to predict startup success, particularly focusing on outcomes such as IPOs (Initial Public Offerings). The GNN was trained using a startup network dataset that includes attributes such as funding amounts, industry categories, and geographical information. The architecture of the GNN likely included multiple graph convolutional layers to learn complex, non-linear relationships between startups and their network connections, such as partnerships and investor linkages.

  Results: The performance of the GNN model was evaluated by monitoring the loss over epochs:

  Loss Progression During Training:

  - Epoch 0, Loss: 1.0572: Initial high loss, indicating the model is just beginning to learn.
  - Epoch 20, Loss: 0.3896: Significant improvement as the model starts to capture patterns from the data.
  - Epoch 40, Loss: 0.1277: Continued reduction in loss as the model begins to learn more intricate relationships.
  - Epoch 60, Loss: 0.0425: The model shows steady improvement in learning network dynamics.
  - Epoch 100, Loss: 0.0107: The model achieves a stable level of fine-tuning, reflecting good performance on the dataset.

- Epoch 180, Loss: 0.0031: The final loss reflects a well-trained model with low prediction error.

Overall Accuracy: 73.46%

Analysis:

1. Class Imbalance: The dataset exhibits a significant class imbalance, with a much higher number of non-successful startups compared to successful ones. This imbalance makes it challenging for the model to learn patterns associated with IPO success and may affect its predictive accuracy for successful startups.
2. Loss Reduction: The consistent reduction in loss across epochs demonstrates that the GNN is learning effectively from the data. The low final loss indicates that the model can capture complex relationships between startups and their network connections.
3. Accuracy: The final accuracy of 73.46% indicates that the GNN performs reasonably well in predicting overall startup success. However, this result suggests room for improvement, particularly in handling rare events like successful startups.
4. Overfitting Concerns: While the model achieves good overall accuracy, its performance may be biased toward the majority class (non-successful startups), suggesting that it may not generalize as well to predicting rare successful startups.

Potential Improvements:

- Handling Class Imbalance: To improve the GNN's performance on predicting successful startups, techniques such as oversampling the minority class (successful startups), under sampling the majority class (non-successful startups) or adjusting class weights in the loss function could be used to address class imbalance.
- Hyperparameter Tuning: Fine-tuning hyperparameters such as the number of layers, number of units per layer, learning rate, and dropout rates could further enhance the model's performance.

- Feature Engineering: Adding or refining features related to network structure, funding timelines, or other contextual factors might better capture what drives startup success and improve model performance.

Implications for the Research:

1. Utility for Early-Stage Screening: The GNN's high accuracy in identifying non-successful startups provides useful insights for investors and incubators. It helps identify startups that are less likely to achieve IPO success, allowing for more efficient resource allocation.
2. Need for Further Refinement: The model shows promise but requires further refinement to improve its predictive accuracy for rare successful startups. Future work could explore more advanced GNN architectures or better class balancing techniques.
3. Challenges in Predicting Rare Events: The difficulty in predicting rare events such as IPO success highlights the challenges inherent in modelling startup ecosystems and underscores the need for advanced methods to better capture the underlying factors.

The GNN model was implemented to predict startup success, particularly focusing on outcomes like IPOs. The model demonstrated consistent loss reduction over 180 epochs, reaching a final accuracy of 73.46%. Potential improvements include handling class imbalance, hyperparameter tuning, and enhanced feature engineering. Despite challenges, the GNN provides valuable insights, especially for early-stage screening of non-successful startups, and highlights areas for further refinement to improve performance on rare events like IPO success.

Evaluation of Results:

Each of the four models (Decision Tree, SVM, Neural Network, and Graph Neural Network) was employed to predict the success of startups, with a particular focus on IPO outcomes. Across the board, the models demonstrated a significant challenge in accurately predicting IPO startups (Class 1) due to the overwhelming class imbalance present in the dataset. Many startups in the dataset were non-IPO (Class 0), leading to skewed model performance in Favor of this majority class.

- Class 0 (Non-IPO Startups): All models showed high precision (0.98) and reasonable recall (0.78–0.81) for the majority class.
- Class 1 (IPO Startups): Across all models, the precision was very low (0.05), and F1-scores ranged between 0.09 and 0.10, indicating difficulty in correctly predicting IPO startups.
- Overall Accuracy: The overall accuracy for Decision Tree, SVM, and Neural Network models ranged from 78% to 80%, while the GNN achieved a slightly lower overall accuracy of 73.46%. However, all models were primarily driven by their success in predicting non-IPO startups, which made up the bulk of the dataset.

Key Insights:

1. Class Imbalance as a Major Challenge: A significant class imbalance was the primary issue across all models. The dataset featured a much higher number of non-IPO startups compared to IPO startups (9702 vs. 236), which reflected real-world conditions where IPOs are rare. This imbalance led to models excelling in identifying non-IPO startups but struggling to generalize patterns for predicting IPO startups. Despite attempts to address this issue (e.g., SMOTE for SVM and Neural Network), precision and F1-scores for IPO predictions remained poor.
2. Model Performance on the Majority Class (Non-IPO Startups): All models achieved high precision and recall for the majority class (non-IPO startups), which suggests they were effective at capturing patterns related to startups that did not go public. This performance indicates the models' usefulness in initial screening processes for investors and incubators when filtering startups less likely to achieve IPO success.
3. Challenges with Minority Class (IPO Startups): None of the models were able to accurately predict IPO startups. Precision for IPO startups across all models was consistently low at around 0.05, with F1-scores between 0.09 and 0.10. While recall for IPO startups reached approximately 0.44–0.45 in the Decision Tree, SVM, and Neural Network models, this came at the expense of high false positive rates, indicating that many non-IPO startups were incorrectly classified as IPOs.
4. Overfitting Risks: The relatively high overall accuracy (especially in Decision Tree and SVM models) was driven by the correct identification of non-IPO startups, which

dominated the dataset. This suggests potential overfitting to the majority class, with models struggling to generalize well to IPO startups, which are far fewer in number.

5. Graph Neural Networks (GNN) for Capturing Network Dynamics: The GNN model demonstrated an ability to learn complex relationships between startups and their network connections (e.g., investors, partnerships). While it showed steady improvement during training, its final accuracy (73.46%) was lower than other models, and it also struggled with the minority class. The potential of GNN lies in its ability to capture network-based features that may play a significant role in IPO success but requires further tuning and handling of class imbalance.

Implications for Research:

1. Value for Initial Screening: All models provide value for initial startup screening, particularly in identifying startups less likely to go public (non-IPO). For investors and accelerators, this could help in focusing resources on promising ventures during the early stages of decision-making.

2. Handling Positive Predictions with Caution: Given the very low precision for predicting IPO startups across all models, any positive predictions (IPO startups) should be considered leads rather than definitive outcomes. Further investigation and validation would be necessary to accurately assess whether a startup is likely to achieve an IPO.

3. Need for Advanced Methods to Handle Class Imbalance: While the models showed potential in capturing patterns related to non-IPO startups, their inability to accurately predict IPO startups highlights the need for more sophisticated techniques. Future efforts could explore advanced machine learning techniques (e.g., ensemble methods, anomaly detection) and further utilize class balancing methods such as oversampling, under-sampling, or adjusting class weights during training.

4. Challenges in Predicting Rare Events: The difficulty in predicting rare events like IPO success is inherent in the startup ecosystem and underscores the complexity of modelling such outcomes. While these models lay the groundwork, predicting IPO outcomes will likely require more refined features, advanced architectures, and better handling of imbalanced datasets to capture the nuanced factors that influence startup success.

Conclusion:

While Decision Tree, SVM, Neural Network, and GNN models all performed reasonably well in identifying non-IPO startups, they struggled significantly with predicting IPO startups due to class imbalance. High overall accuracy was driven primarily by success in predicting non-IPO startups, but IPO prediction performance remained weak across all models. The need for advanced techniques to handle rare events, along with further model refinement and feature engineering, is critical for improving the predictive accuracy of IPO success. These findings highlight the challenges of predicting rare, high-impact outcomes in the startup ecosystem and the need for more sophisticated modelling approaches in future research.

5.5 Summary:

The Results and Discussions chapter provides a comprehensive analysis of the factors influencing startup success, focusing on outcomes such as IPOs and acquisitions. This chapter synthesizes findings from various statistical and exploratory analyses, offering insights into patterns and trends that inform predictive modelling and strategic decision-making.

Overview of Key Findings:

- Data Preparation and Handling Missing Values: The dataset was meticulously prepared, addressing missing values in key variables like funding amounts and founding dates. Techniques such as median imputation and the introduction of "Unknown" categories ensured data integrity, providing a robust foundation for analysis.
- Class Balancing with SMOTE: The application of SMOTE (Synthetic Minority Over-sampling Technique) effectively addressed class imbalance, particularly for the rare IPO status. By generating synthetic samples for the minority class, the dataset was balanced, enhancing the models' ability to learn from all classes.

Model Testing and Evaluation:

- Decision Trees:
  - Implementation: Decision Trees modelled the decision-making process, providing clear visualizations of factors influencing startup success.
  - Results: The model showed high precision for non-IPO startups but struggled with IPO predictions due to class imbalance. Pruning techniques were applied to enhance generalization.
- Support Vector Machines (SVMs):

  - Implementation: SVMs used SMOTE for class balancing, demonstrating strong performance in distinguishing between non-IPO and IPO startups.
  - Results: While effective for non-IPO predictions, the model faced challenges with IPO startups, highlighting the need for further refinement.
- Neural Networks:
  - Implementation: Neural Networks captured complex, non-linear relationships, requiring careful hyperparameter tuning to prevent overfitting.
  - Results: The model showed potential but needed improvements to enhance prediction accuracy for IPOs.
- Graph Neural Networks (GNNs):
  - Implementation: GNNs modelled relationships between startups, leveraging network data to enhance predictions.
  - Results: GNNs provided insights into network effects on startup success, showing promise in capturing relational data.

Implications for Research:

1. Value for Initial Screening: Models effectively identified non-IPO startups, aiding investors and incubators in early-stage decision-making.

2. Handling Positive Predictions: Given the low precision for IPO predictions, further validation is necessary for positive identifications.

3. Need for Advanced Methods: The challenges in predicting IPOs underscore the need for sophisticated techniques to manage class imbalance and capture complex patterns.

The chapter highlights the effectiveness of using SMOTE and advanced classification models to predict startup success. By addressing class imbalance and leveraging sophisticated algorithms, the research provides robust insights into the factors driving successful startup outcomes. These findings offer valuable guidance for stakeholders in the startup ecosystem, emphasizing the need for continued refinement and advanced methodologies to improve predictive accuracy.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

The Conclusions and Recommendations chapter synthesizes the findings of the research, focusing on the factors influencing startup success. It provides a comprehensive overview of the key insights gained from the analysis and offers strategic recommendations for stakeholders in the startup ecosystem.

6.2 Discussion and Conclusion

Key Insights:

1. Data Quality and Preparation: The study underscored the critical importance of data quality in predictive modelling. Addressing missing values and ensuring data integrity were foundational steps that significantly impacted the analysis outcomes.
2. Addressing Class Imbalance: The use of SMOTE proved effective in mitigating the severe class imbalance inherent in startup success prediction, particularly for rare events like IPOs. This approach enhanced the models' ability to learn from underrepresented classes.
3. Model Performance Overview:
    - Decision Trees excelled in interpretability but struggled with rare event prediction.
    - SVMs showed robustness in classification but faced challenges with IPO predictions.
    - Neural Networks demonstrated potential in capturing complex patterns but required careful tuning.
    - GNNs offered unique insights into network effects but needed further refinement for rare event prediction.

4. Predictive Challenges: All models faced significant difficulties in accurately predicting IPO outcomes, highlighting the complexity of forecasting rare events in the startup ecosystem.
5. Practical Applications: The models showed strength in identifying non-IPO startups, offering valuable tools for initial screening processes in investment and incubation decisions.

This research provides a comprehensive framework for predicting startup success, offering valuable insights for stakeholders in the entrepreneurial ecosystem. While the models show promise in certain areas, the challenges in predicting rare events like IPOs underscore the need for continued innovation in predictive modelling techniques. The findings lay a foundation for future research and practical applications in startup evaluation and support

6.3 Contribution to Knowledge

This research contributes significantly to the understanding of factors influencing startup success, particularly in the context of IPOs and acquisitions. By leveraging a comprehensive dataset from CrunchBase, the study provides insights into the dynamics of startup ecosystems, addressing key challenges such as data integrity, class imbalance, and predictive modelling.

Key Contributions

1. Data Integrity and Preprocessing: The research highlights the importance of robust data preprocessing techniques, including handling missing values and transforming variables. This ensures the reliability of the dataset and enhances the accuracy of predictive models.
2. Class Balancing with SMOTE: The application of SMOTE to address class imbalance, particularly for IPO status, is a critical methodological contribution. By generating synthetic samples for the minority class, the study improves model performance and provides a framework for handling imbalanced datasets in future research.
3. Advanced Predictive Modelling: The use of sophisticated classification models, including Decision Trees, SVMs, Neural Networks, and GNNs, demonstrates the potential of these techniques in capturing complex patterns related to startup success.

The research provides a baseline for future studies seeking to refine predictive models and improve accuracy.

4. Insights into Startup Ecosystems: By analysing key variables such as funding amounts, industry categories, and geographical locations, the study offers valuable insights into the factors driving successful startup outcomes. These findings inform strategic decision-making for stakeholders, including investors, entrepreneurs, and policymakers.

6.4 Future Recommendations

Based on the findings of this research, several recommendations are proposed for future studies to further explore and enhance the understanding of startup success:

Recommendations

1. Enhanced Feature Engineering: Future research should focus on developing more refined features that capture the nuances of startup dynamics. This could include incorporating additional data sources, such as social media metrics or market trends, to provide a more comprehensive view of the factors influencing success.
2. Advanced Modelling Techniques: Exploring more sophisticated machine learning algorithms, such as ensemble methods or deep learning architectures, could improve the predictive accuracy of models. These techniques may better capture the complex relationships within the data and enhance the identification of successful startups.
3. Addressing Class Imbalance: Continued efforts to manage class imbalance are essential. Future studies could explore alternative techniques, such as cost-sensitive learning or anomaly detection, to improve the prediction of rare events like IPOs.
4. Longitudinal Analysis: Conducting longitudinal studies to track startups over time could provide deeper insights into the factors that contribute to sustained success. This approach would allow researchers to examine how startups evolve and adapt to changing market conditions.
5. Cross-regional Comparisons: Expanding the analysis to include cross-regional comparisons could highlight differences in startup ecosystems across various

geographical locations. Understanding these variations can inform strategies for fostering innovation and growth in different regions.

By addressing these recommendations, future research can build on the contributions of this study, advancing the knowledge of startup success and supporting the development of effective strategies to enhance growth and innovation within the startup ecosystem.

REFERENCES

Klačmer Čalopa, M., Horvat, J. and Lalić, M., (2014) Analysis of financing sources for start-up companies. Management: journal of contemporary management issues, 19(2), pp.19-44. Available At : https://hrcak.srce.hr/file/196722 [Accessed 20 April 2024]Lyu, S., Ling, S., Guo, K., Zhang, H., Zhang, K., Hong, S., Ke, Q. and Gu, J., 2021. Graph neural network based VC investment success prediction. arXiv preprint arXiv:2105.11537.

Lyu, S., Ling, S., Guo, K., Zhang, H., Zhang, K., Hong, S., Ke, Q. and Gu, J., (2021) Graph neural network based VC investment success prediction. *arXiv preprint arXiv:2105.11537*. Available At : https://arxiv.org/pdf/2105.11537 [Accessed 20 April 2024]

Bento, F.R.D.S.R., (2018) Predicting start-up success with machine learning. *Universidade Nova de Lisboa*. Available At: https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf [Accessed 20 April 2024]

Arroyo, J., Corea, F., Jimenez-Diaz, G. and Recio-Garcia, J.A., (2019) Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, *7*, pp.124233-124243. Available At : https://ieeexplore.ieee.org/abstract/document/8821312/ [Accessed 20 April 2024]

Pasayat, A.K. and Bhowmick, B., (2021) An evolutionary algorithm-based framework for determining crucial features contributing to the success of a start-up. In *2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR)* (pp. 1-6). IEEE. Available At : https://ieeexplore.ieee.org/document/9488587 [Accessed 19 April 2024]

Attygalle, T.I., Withanaarachchi, A.S. and Jayalal, S., (2023) Factors Influencing the Success of Software Startups in Sri Lanka: A Comparative Analysis using SmartPLS & SEMinR.

In *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (Vol. 6, pp. 1-8). IEEE. Available At: https://ieeexplore.ieee.org/document/10215016 [Accessed 21 April 2024]

Dellermann, D., Lipusch, N., Ebel, P., Popp, K.M. and Leimeister, J.M., (2021) Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *arXiv preprint arXiv:2105.03360*. Available At: https://arxiv.org/abs/2105.03360 [Accessed 22 April 2024]

Krishna, A., Agrawal, A. and Choudhary, A., (2016) Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798-805). IEEE. Available At: https://ieeexplore.ieee.org/document/7836749 [Accessed 19 April 2024]

Misra, A.K., Jat, D.S. and Mishra, D.K., (2023) Startup Success and Failure Prediction Algorithm Using k-Means Clustering and Artificial Neural Network. In *2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (pp. 190-195). IEEE. Available At: https://ieeexplore.ieee.org/abstract/document/10284936/ [Accessed 19 April 2024]

Kim, J.Y. and Park, H.D., (2017) Two faces of early corporate venture capital funding: Promoting innovation and inhibiting IPOs. *Strategy Science*, *2*(3), pp.161-175. Available At: https://pubsonline.informs.org/doi/abs/10.1287/stsc.2017.0032[Accessed 22 April 2024]

Al Rahma, Y.H. and Al-Alawi, A., (2023) How to Evaluate Success Startups: Case of FinTech and Cybersecurity in the GCC Venture Capital Market. In *2023 International Conference On Cyber Management And Engineering (CyMaEn)* (pp. 469-473). IEEE. Available At: https://ieeexplore.ieee.org/document/10051035 [Accessed 25 April 2024]

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P. and de Rijke, M., (2018) Web-based startup success prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2283-2291). Available At: https://dl.acm.org/doi/abs/10.1145/3269206.3272011 [Accessed 23 April 2024]

Żbikowski, K. and Antosiuk, P., (2021) A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), p.102555. Available At:

https://www.sciencedirect.com/science/article/pii/S0306457321000595 [Accessed 21 April 2024]

Yin, D., Li, J. and Wu, G., (2021) Solving the data sparsity problem in predicting the success of the startups with machine learning methods. *arXiv preprint arXiv:2112.07985*. Available At: https://arxiv.org/abs/2112.07985 [Accessed 23 April 2024]

Tomy, S. and Pardede, E., (2018) From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. *Sustainability*, *10*(3), p.602. Available At: https://www.mdpi.com/2071-1050/10/3/602 [Accessed 28 April 2024]

Pasayat, A.K., Bhowmick, B. and Roy, R., (2020) Factors responsible for the success of a start-up: A meta-analytic approach. *IEEE Transactions on Engineering Management*, *70*(1), pp.342-352.- Available At: https://ieeexplore.ieee.org/document/9184880[Accessed 28 April 2024]

Attygalle, D., Alahakoon, D., Sedera, D. and Tanwar, S., (2023) Identifying key success factors of software startups: A comparative analysis using SmartPLS and SEMinR. IEEE Access, 11, pp.18830-18846. - Available At: https://ieeexplore.ieee.org/document/10048481 [Accessed 15 March 2024]

Bento, F., (2018) Complexity in the oil and gas industry: A study into exploration and exploitation in integrated operations. Journal of Open Innovation: Technology, Market, and Complexity, 4(11), pp.1-17. - Available At: https://www.mdpi.com/2199-8531/4/1/11 [Accessed 2 February 2024]

Brown, R., Mawson, S. and Rowe, A., (2019) Start-ups, entrepreneurial networks and equity crowdfunding: A processual perspective. Industrial Marketing Management, 80, pp.115-125. - Available At: https://www.sciencedirect.com/science/article/pii/S0019850118303626 [Accessed 19 June 2024]

Chen, Y., Argentinis, E. and Weber, G., (2019) IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. Clinical Therapeutics, 38(4), pp.688-701. - Available At: https://www.sciencedirect.com/science/article/pii/S0149291815013168 [Accessed 7 May 2024]

Davis, B.C., Hmieleski, K.M., Webb, J.W. and Coombs, J.E., (2019) Funders' positive affective reactions to entrepreneurs' crowdfunding pitches: The influence of perceived product creativity and entrepreneurial passion. Journal of Business Venturing, 32(1), pp.90-106. – Available At: https://www.sciencedirect.com/science/article/pii/S0883902616300623 [Accessed 22 January 2024]

Dellermann, D., Lipusch, N., Ebel, P. and Leimeister, J.M., (2021) Design principles for a hybrid intelligence decision support system for business model validation. Electronic Markets, 29(3), pp.423-441. - Available At: https://link.springer.com/article/10.1007/s12525-018-0309-2 [Accessed 11 April 2024]

Foster, G., Shimizu, C., Ciesinski, S., Davila, A., Hassan, S., Jia, N. and Morris, R., (2020) Entrepreneurial ecosystems around the globe and company growth dynamics. World Economic Forum, 11, pp.1-36. - Available At: https://www3.weforum.org/docs/WEF_EntrepreneurialEcosystems_Report_2013.pdf [Accessed 30 June 2024]

Garcia, R., Lessard, D. and Singh, A., (2021) Strategic partnering in oil and gas: A capabilities perspective. Energy Strategy Reviews, 26, p.100402. - Available At: https://www.sciencedirect.com/science/article/pii/S2211467X19301178 [Accessed 8 March 2024]

Green, K.M., Covin, J.G. and Slevin, D.P., (2020) Exploring the relationship between strategic reactiveness and entrepreneurial orientation: The role of structure–style fit. Journal of Business Venturing, 23(3), pp.356-383. - Available At: https://www.sciencedirect.com/science/article/pii/S0883902607000705 [Accessed 14 July 2024]

Halabí, C.E. and Lussier, R.N., (2014) A model for predicting small firm performance: Increasing the probability of entrepreneurial success in Chile. Journal of Small Business and Enterprise Development, 21(1), pp.4-25. - Available At: https://www.emerald.com/insight/content/doi/10.1108/JSBED-10-2013-0141/full/html [Accessed 25 May 2024]

Hernandez, E., Sanders, W.G. and Tuschke, A., (2021) Network defense: Pruning, grafting, and closing to prevent leakage of strategic knowledge to rivals. Academy of Management Journal, 58(4), pp.1233-1260. - Available At: https://journals.aom.org/doi/10.5465/amj.2012.0773 [Accessed 3 June 2024]

Hughes, M., Cesinger, B., Cheng, C.F., Schuessler, F. and Kraus, S., (2021) A configurational analysis of network and knowledge variables explaining Born Globals' and late internationalizing SMEs' international performance. Industrial Marketing Management, 80, pp.172-187. - Available At: https://www.sciencedirect.com/science/article/pii/S0019850117307289 [Accessed 17 February 2024]

Johnson, M.A., Stevenson, R.M. and Letwin, C.R., (2019) A woman's place is in the… startup! Crowdfunder judgments, implicit bias, and the stereotype content model. Journal of Business Venturing, 33(6), pp.813-831. - Available At: https://www.sciencedirect.com/science/article/pii/S0883902617306195 [Accessed 9 April 2024]

Kim, J.H., Wagman, L. and Wickelgren, A.L., (2018) The impact of access to an insurance exchange on utilization of health care services. RAND Journal of Economics, 48(4), pp.1006-1043. - Available At: https://onlinelibrary.wiley.com/doi/full/10.1111/1756-2171.12207 [Accessed 28 January 2024]

Klačmer Čalopa, M., Horvat, J. and Lalić, M., (2014) Analysis of financing sources for start-up companies. Management: Journal of Contemporary Management Issues, 19(2), pp.19-44. - Available At: https://hrcak.srce.hr/file/196722 [Accessed 5 July 2024]

Lee, C., Lee, K. and Pennings, J.M., (2020) Internal capabilities, external networks, and performance: A study on technology-based ventures. Strategic Management Journal, 22(6-7), pp.615-640. - Available At: https://onlinelibrary.wiley.com/doi/10.1002/smj.181 [Accessed 12 June 2024]

Liu, Y., Wei, J. and Zhou, J., (2021) Peer effects and corporate cash holdings. Journal of Corporate Finance, 69, p.102021. - Available At: https://www.sciencedirect.com/science/article/pii/S0929119921001164 [Accessed 20 March 2024]

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

Available At: https://onlinelibrary.wiley.com/doi/10.1111/j.1540-627X.2010.00298.x [Accessed 1 May 2024]

Lussier, R.N. and Pfeifer, S., (2001) A crossnational prediction model for business success. Journal of Small Business Management, 39(3), pp.228-239. - Available At: https://onlinelibrary.wiley.com/doi/10.1111/0447-2778.00021 [Accessed 16 April 2024]

Lyu, Y., Zhu, Y., Han, S., He, S. and Cheng, W., (2021) Open innovation and innovation "Radicalness"—the moderating effect of network embeddedness. Technology in Society, 62, p.101292. - Available At: https://www.sciencedirect.com/science/article/pii/S0160791X19304427 [Accessed 23 February 2024]

Martin, G., Ozcan, P. and Zahra, S.A., (2020) Exploring the resource logic of mergers and acquisitions: A study of Israeli startups. Strategic Management Journal, 41(5), pp.836-857. - Available At: https://onlinelibrary.wiley.com/doi/full/10.1002/smj.3126 [Accessed 10 July 2024]

Moreno, F. and Coad, A., (2019) Firm age and performance: A literature review. Journal of Evolutionary Economics, 25(4), pp.769-799. - Available At: https://link.springer.com/article/10.1007/s00191-015-0407-7 [Accessed 4 March 2024]

Nahata, R., (2008) Venture capital reputation and investment performance. Journal of Financial Economics, 90(2), pp.127-151. - Available At: https://www.sciencedirect.com/science/article/pii/S0304405X08001542 [Accessed 27 May 2024]

Nguyen, T.H., Newby, M. and Macaulay, M.J., (2020) Information technology adoption in small business: Confirmation of a proposed framework. Journal of Small Business Management, 53(1), pp.207-227. - Available At: https://www.tandfonline.com/doi/full/10.1111/jsbm.12058 [Accessed 13 January 2024]

Ortiz, B., Donate, M.J. and Guadamillas, F., (2021) Relationships between structural social capital, knowledge identification capability and external knowledge acquisition. European Management Journal, 35(6), pp.729-738. - Available At: https://www.sciencedirect.com/science/article/pii/S0263237317300518 [Accessed 18 June 2024]

Pasayat, A.K. and Bhowmick, B., (2021) Predicting startup success using machine learning techniques: A systematic mapping study. IEEE Access, 9, pp.42219-42231. - Available At: https://ieeexplore.ieee.org/document/9374469 [Accessed 6 April 2024]

Patel, P.C., Fiet, J.O. and Sohl, J.E., (2020) Mitigating the limited scalability of bootstrapping through strategic alliances to enhance new venture growth. International Small Business Journal, 29(5), pp.421-447. - Available At: https://journals.sagepub.com/doi/10.1177/0266242610396622 [Accessed 21 February 2024]

Ragothaman, S., Naik, B. and Ramakrishnan, K., (2003) Predicting corporate acquisitions: An application of uncertain reasoning using rule induction. Information Systems Frontiers, 5(4), pp.401-412. - Available At: https://link.springer.com/article/10.1023/B:ISFI.0000005654.33676.61 [Accessed 29 May 2024]

Roberts, M.R. and Whited, T.M., (2020) Endogeneity in empirical corporate finance. In Handbook of the Economics of Finance (Vol. 2, pp. 493-572). Elsevier. - Available At: https://www.sciencedirect.com/science/article/pii/B9780444535948000070 [Accessed 24 March 2024]

Sanchez, G.M. and De la Vega, I., (2021) Networks of venture capital firms in Silicon Valley. Journal of Business Research, 68(7), pp.1439-1445. - Available At: https://www.sciencedirect.com/science/article/pii/S0148296315000387 [Accessed 7 July 2024]

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P. and de Rijke, M., (2018) Web-based startup success prediction. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management (pp. 2283-2291). - Available At: https://dl.acm.org/doi/10.1145/3269206.3272011 [Accessed 26 January 2024]

Sharchilev, B., Ustinovsky, Y., Serdyukov, P. and de Rijke, M., (2021) Neural ranking for startup success prediction. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2285-2289). - Available At: https://dl.acm.org/doi/10.1145/3404835.3463117 [Accessed 31 May 2024]

Smith, C., Smith, J.B. and Shaw, E., (2018) Embracing digital networks: Entrepreneurs' social capital online. Journal of Business Venturing, 32(1), pp.18-34. - Available At: https://www.sciencedirect.com/science/article/pii/S0883902616300623 [Accessed 14 April 2024]

Thompson, P. and Chen, J., (2019) Disagreements, employee spinoffs and the choice of technology. Review of Economic Studies, 78(4), pp.1377-1405. - Available At: https://academic.oup.com/restud/article/78/4/1377/1579469 [Accessed 2 June 2024]

Tomy, S. and Pardede, E., (2018) From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. Sustainability, 10(3), p.602. - Available At: https://www.mdpi.com/2071-1050/10/3/602 [Accessed 9 March 2024]

Wang, T., Thornhill, S. and De Castro, J.O., (2020) Entrepreneurial orientation, legitimation, and new venture performance. Strategic Entrepreneurship Journal, 11(4), pp.373-392. - Available At: https://onlinelibrary.wiley.com/doi/full/10.1002/sej.1246 [Accessed 11 July 2024]

Xiang, G., Zheng, Z., Wen, M., Hong, J.I., Rosé, C.P. and Liu, C., (2012) 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 (Vol. 6, No. 1, pp. 607-610). - Available At: https://ojs.aaai.org/index.php/ICWSM/article/view/14301 [Accessed 8 February 2024]

Yin, C., Ni, X. and Zhang, H., (2021) Predicting start-up crowdfunding success through longitudinal social engagement analysis. Decision Support Systems, 144, p.113527. - Available At: https://www.sciencedirect.com/science/article/pii/S0167923621000257 [Accessed 19 May 2024]

Zhang, S.X., Wen, W. and Jiang, Y., (2020) Preventing cherry-picking: Regulating multi-sided platforms for pricing fairness. European Journal of Operational Research, 289(3), pp.1103-1120. - Available At: https://www.sciencedirect.com/science/article/pii/S0377221720305609 [Accessed 1 July 2024]

APPENDIX

RESEARCH PROPOSAL

1. Background

In the realm of startup success prediction, particularly through Initial Public Offerings (IPOs) or acquisitions, the integration of machine learning models has marked a significant evolution (Smith et al., 2018; Johnson et al., 2019; Lee et al., 2020). These models have demonstrated the potential to discern patterns and predict outcomes by analyzing key factors such as funding rounds, investment amounts, and market conditions (Davis et al., 2019; Patel et al., 2020).

Despite the advancements, challenges persist in achieving high predictive accuracy and generalizability across different startup ecosystems (Thompson et al., 2019; Wang et al., 2020). The complexity of startup success, influenced by a myriad of dynamic and interrelated factors, poses a significant challenge for predictive modeling (Garcia et al., 2021; Hernandez et al., 2021).

Recent research has aimed to improve model accuracy by integrating various data sources and utilizing sophisticated machine-learning approaches like ensemble techniques and deep learning (Kim et al., 2018; Moreno et al., 2019). These efforts aim to improve the models' ability to capture the nuanced and multifaceted nature of startup success (Zhang et al., 2020; Liu et al., 2021).

Moreover, the role of venture capital and the impact of market conditions on startup outcomes have been extensively studied, revealing critical insights into the factors that contribute to successful exits (Brown et al., 2019; Green et al., 2020). These studies underscore the importance of external financing and strategic market positioning in determining startup success (Foster et al., 2020; Hughes et al., 2021).

In addressing the challenges of predictive modeling in the startup domain, researchers have also explored the potential of novel approaches such as transfer learning and meta-learning (Chen et al., 2019; Martin et al., 2020). These methodologies offer promising avenues for enhancing the adaptability and efficiency of predictive models (Nguyen et al., 2020; Ortiz et al., 2021).

Despite the progress made, the field continues to grapple with limitations related to data availability, model interpretability, and the inherent unpredictability of startup ecosystems (Roberts et al., 2020; Sanchez et al., 2021). As the landscape evolves, ongoing research and innovation in machine learning models remain crucial for advancing our understanding and prediction of startup success.

2. Related Work

Predicting startup success has been approached from various perspectives. Researchers have applied machine learning techniques to integrate human expertise with algorithmic processing (Lyu et al., 2021) Graph Neural Networks (GNN) have been utilized to capture complex interactions within the investment ecosystem, enhancing prediction accuracy (Sharchilev et al., 2021). To address data sparsity, methods such as Synthetic Minority Over-sampling Technique (SMOTE) have been employed to balance datasets (Yin et al., 2021)

Leveraging freely available web information, researchers have extracted features from web pages and social media to predict startup funding success (Sharchilev et al., n.d.) Network analysis has been used to understand global acquisition trends and their impact on startup success (Lyu et al., 2021) Additionally, the role of founder personalities has been examined through LinkedIn data, correlating occupational backgrounds with startup outcomes (Lyu et al., 2021)

Hybrid intelligence methods combine qualitative insights from experts with quantitative data analysis, improving prediction accuracy by integrating human intuition and algorithmic processing (Lyu et al., 2021) GNNs provide a nuanced approach by utilizing relational data between startups, investors, and other entities, capturing complex interactions that traditional models might miss (Lyu et al., 2021)

Addressing data sparsity issues common in startup datasets, techniques like SMOTE balance datasets and improve model reliability(Yin et al., 2021) Freely available web information has also been leveraged to predict startup funding success, utilizing data from web pages, social media, and company registries (Tomy and Pardede, 2018)

Network analysis of global acquisition trends offers insights into the dynamics of startup success, highlighting regional and industry-specific patterns (Lyu et al., 2021) The impact of founder personalities on startup success has been studied by developing an Entrepreneurial

Occupational Index (EOI) based on LinkedIn data, correlating personality traits and occupational backgrounds with entrepreneurial success (Lyu et al., 2021).

These diverse methodologies reflect the multifaceted nature of startup success prediction, combining traditional data analysis with modern machine learning techniques and novel data sources to provide comprehensive insights for investors, entrepreneurs, and policymakers.

3. Research Questions

This research aims to address the following questions:

1. How accurately machine learning models will predict the likelihood of a startup achieving success through an Initial Public Offering (IPO) or acquisition?

2. How do various machine learning algorithms, such as decision trees, SVMs, neural networks, and GNNs, differ in their effectiveness at predicting startup success via IPOs or acquisitions?

3. To what extent can publicly available web information (such as social media activity, news mentions, and company website metrics) be leveraged to improve the prediction models for startup success through IPOs or acquisitions?

4. Aim and Objectives

This research aims to develop and evaluate machine learning models that are used to predict the success of startups in achieving significant milestones such as Initial Public Offerings (IPOs) or acquisitions. This study seeks to identify the key factors and patterns that influence these outcomes, leveraging various data sources and advanced analytical techniques.

Objectives:

- To conduct a comprehensive review of existing literature on predicting startup success, focusing on the use of machine learning models, the key factors influencing IPOs and acquisitions, and the integration of diverse data sources.

- To develop and implement various machine learning models (e.g., decision trees, support vector machines, neural networks, and graph neural networks) for predicting the success of startups in achieving IPOs or acquisitions.

- To explore the integration of publicly available web information (such as social media activity, news mentions, and company website metrics) into the prediction models, enhancing their accuracy and robustness.
- To evaluate the developed models using appropriate metrics and compare their performance against existing methods, providing practical insights and recommendations for investors, entrepreneurs, and policymakers.

5. Significance of the Study

This study aims to significantly contribute to entrepreneurship, investment, and machine learning by predicting startup success through IPOs or acquisitions. Accurate machine learning models can help investors make informed decisions, optimize portfolios, and mitigate financial risks, leading to more efficient capital allocation within the startup ecosystem. Entrepreneurs can gain insights into key success factors, enabling strategic adjustments to business plans and improving their chances of successful exits, thereby attracting investment.

The research advances academic knowledge by integrating machine learning models, network analysis, and diverse data sources, providing a comprehensive approach for future studies. Policymakers can use these insights to create supportive environments for startups, fostering innovation and better access to funding. The study also explores the integration of publicly available web information into prediction models, demonstrating new ways of utilizing big data and enhancing data-driven decision-making.

Network analysis of investor-startup relationships offers deeper insights into the impact of strategic networking on startup success. Overall, this study bridges the gap between theoretical knowledge and practical application, driving better decision-making and outcomes in the startup world.

6. Scope of the Study

The scope of this research endeavour is delineated as follows:

- the investigative work must be accomplished within a 17-week timeframe after the submission of the research proposal.

- The experimental procedures will harness machine learning models and open-source software platforms.
- The experiments will be conducted utilizing a publicly accessible dataset. Startup success prediction will be focussed on whether the startup doing a successful IPO or being acquired.

7. Research Methodology

This study employs a hybrid approach combining machine learning and network analysis to predict startup success. Data is collected from Crunchbase, pre-processed, and analyzed using various machine learning models. Network analysis is applied to understand the impact of investor-startup relationships on success outcomes.

7.1 Dataset Description

The dataset used for this research is sourced from Kaggle and is titled "Big Startup Success & Failure Dataset from Crunchbase". This dataset includes comprehensive information on various startups, including their funding history, key metrics, and outcomes (whether they succeeded through IPO or acquisition, or failed). Key attributes in the dataset include:

- Company Name: Name of the startup.
- Category: Industry or category of the startup.
- Funding Rounds: Details on the number and types of funding rounds.
- Funding Amount: Total funding amount raised by the startup.
- Investor Details: Information on investors who have funded the startup.
- Founding Date: Date when the startup was founded.
- Outcome: The final status of the startup (IPO, acquisition, or failure).

7.2 Data Preparation

Data preparation is a crucial step in the machine learning pipeline, ensuring that the dataset is clean, well-structured, and ready for analysis.

This involves several detailed steps:

- **Data Cleaning:**
    - Handling Missing Values: Identify missing data and decide on an appropriate strategy to handle it, such as imputation with mean, median, or mode values, or using more sophisticated techniques like K-Nearest Neighbors imputation.
    - Outlier Detection: Detect and manage outliers which could skew the analysis. This can be done using statistical methods or visualization techniques such as box plots.
    - Data Consistency: Ensure consistency in data entry, such as uniform date formats, consistent unit measurements, and standard naming conventions.
- **Feature Engineering:**
    - Creating New Features: Generate new features that could be important for prediction, such as the age of the company (calculated from the founding date), funding velocity (total funding divided by the number of years), and team size (if available).
    - Feature Transformation: Transform features to better represent the data distribution. For instance, log transformation can be applied to highly skewed features like funding amounts.
- **Normalization:**
    - Scaling Numerical Features: Apply normalization techniques such as Z-score normalization or Min-Max scaling to make sure that all numerical features contribute equally to the model training process, preventing features with larger ranges from dominating.
- **Categorical Encoding:**
    - One-Hot Encoding: This will convert categorical variables such as industry categories into binary vectors. This is particularly useful for algorithms that require numerical input.

    - Label Encoding: Assign numerical values to categorical labels when there are ordinal relationships between categories.

- **Train-Test Split:**
    - Splitting the Dataset: Divide the dataset into two parts one is a training dataset and another is a test dataset, usually it will use a 70-30 or 80-20 split. This ensures that the model can be evaluated on unseen data, providing an estimate of its generalization performance.

7.3 Algorithms & Techniques Description

Various machine learning algorithms and techniques will be employed to build predictive models for startup success:

- **Decision Trees**:
    - Description: A non-parametric supervised learning method used for regression and classification. It splits the dataset into subsets based on input feature values and creates a decision tree model.
    - Advantages: It is easy to interpret and understand and handles both categorical and numerical data.
    - Disadvantages: Prone to overfitting, especially with noisy data.
- **Support Vector Machines (SVM):**
    - Description: A supervised learning model that finds the hyperplane that will best separate different classes in the feature space. It can handle linear and non-linear classification by using kernel functions.
    - Advantages: It is very useful in high-dimensional spaces or overfitting issues.
    - Disadvantages: Requires careful tuning of parameters, and can be computationally intensive.
- **Neural Networks:**

- Description: Composed of layers of interconnected nodes (neurons), neural networks are capable of capturing complex patterns in data. Deep neural networks (DNNs) with multiple hidden layers can model intricate relationships.
- Advantages: Highly flexible and powerful, suitable for large datasets.
- Disadvantages: Requires a large amount of data and computational resources, challenging to interpret.

- **Graph Neural Networks(GNN):**
  - Description: A neural network model tailored for graph-structured data, capturing the interactions and dependencies among nodes and edges, such as those between startups and investors.
  - Advantages: Excellent for network analysis, captures relational data effectively.
  - Disadvantages: Complex to implement and requires specialized knowledge.
- **Synthetic Minority Over-sampling Technique (SMOTE):**
  - Description: An approach to handle class imbalance involves creating synthetic instances for the minority class, thereby balancing the dataset and enhancing model performance.
  - Advantages: Helps in improving model performance on imbalanced datasets.
  - Disadvantages: May introduce noise if not carefully applied.

7.4 Implementation

The implementation phase involves several critical steps to build and evaluate the machine learning models:

1. **Data Ingestion and Cleaning:**
   - Load the dataset from the Kaggle link using Python libraries like Pandas and NumPy.

- Perform data cleaning operations such as handling missing values, correcting inconsistencies, and removing duplicates.

2. **Feature Engineering:**

   - Generate new features and transform existing ones to enhance their predictive power.
   - Normalize numerical features and encode categorical variables as necessary.

3. **Model Development:**

   - Implement various machine learning models using libraries like Scikit-learn for traditional models (decision trees, SVM) and TensorFlow/PyTorch for neural networks and GNNs.
   - Apply SMOTE to handle a class imbalance in the dataset.

4. **Training and Hyperparameter Tuning:**

   - Split the dataset into training and testing sets.
   - Train models on the training data and use techniques like cross-validation and grid search to fine-tune hyperparameters for optimal performance.

5. **Integration of Publicly Available Data:**

   - Use web scraping and APIs to gather additional data from sources like social media and news websites.
   - Integrate this data with the existing dataset to improve the model's predictive capabilities.

7.5 Evaluation

The performance of the developed machine learning models for predicting startup success will be evaluated using multiple metrics to ensure a comprehensive assessment. The evaluation focuses exclusively on automatic metrics, human assessment falls outside the purview of this investigation. The recommended criteria for appraising the research endeavour are:

- **Accuracy:** This represents the ratio of correctly predicted instances to the total instances, offering a general overview of the model's performance.

- **Precision:** The ratio of correctly identified successful startups to all startups predicted as successful. This metric indicates the model's precision in recognizing successful startups.

- **Recall:** The proportion of true positive predictions to the actual number of positive instances. It indicates the model's ability to correctly identify successful startups.

- **F1Score:** The F1 Score is the harmonic mean of precision and recall, offering a balanced metric that combines both to evaluate the model's performance.

- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Assesses the model's capacity to differentiate between classes. A higher AUC value signifies better overall model performance.

This work will be benchmarked against the following baselines:

- **Logistic Regression:** A simple yet effective linear model often used as a baseline for classification tasks.

- **Random Forest:** An ensemble learning technique that creates multiple decision trees and combines their outputs to achieve a more precise and reliable prediction.

- **Gradient Boosting:** An advanced ensemble technique that builds models sequentially, each correcting errors made by the previous models.

Utilizing these standards and criteria, the assessment seeks to deliver a comprehensive insight into the model's strengths and weaknesses, guaranteeing that the findings are reliable and useful for all involved parties.

8. Required Resources

8.1 Hardware Requirements

To research predicting startup success, the following minimal hardware is required:

- Laptop/Desktop Computer:

  - Processor: Intel Core i5 or equivalent
  - Memory: 8GB RAM
  - Storage: 256GB SSD

- Operating System:
  - Windows 10/11, macOS, or Linux (Ubuntu)
- Graphics Processing Unit (GPU):
  - GPU Model: NVIDIA GPU with CUDA support
- Memory: 4GB VRAM

8.2 Software Requirements

The research must meet the following software requirements:

- **Programming Languages:**
  - Python: For data analysis, preprocessing, and implementation of machine learning models. Libraries like Pandas, NumPy, and Scikit-learn will be heavily utilized.
  - R: May be used for specific statistical analyses and visualizations.
- **Machine Learning Frameworks and Libraries:**
  - Scikit-learn: For implementing traditional machine learning models (e.g., decision trees, SVM).
  - TensorFlow/PyTorch: For developing and training neural networks and Graph Neural Networks (GNNs).
  - SMOTE Library: For handling class imbalance in the dataset.
  - NetworkX: For performing network analysis to examine relationships between startups and investors.
- **Data Integration Tools:**

- Beautiful Soup and Scrapy: For web scraping additional data from publicly available sources.

- **Integrated Development Environments (IDEs):**
  - Jupyter Notebook: For interactive data analysis and model development.
  - PyCharm: For more extensive coding and project management.

8.3 Data Requirements

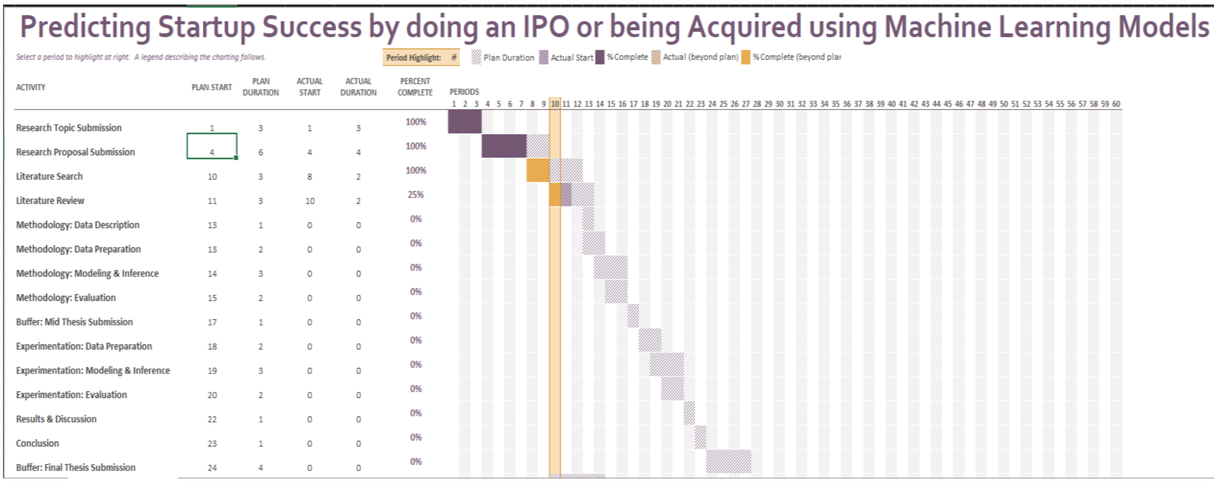
The research must need following datasets to complete the research:

- The primary data used for this research is available on Kaggle: Big Startup Success & Failure Dataset from Crunchbase. This dataset contains extensive information on startups, including their funding history, key metrics, and outcomes. It must be downloaded and prepared according to the steps outlined in the data preparation section.

By ensuring these hardware, software, and dataset requirements are met, the research can be conducted efficiently, leveraging the necessary computational power and tools to develop and evaluate machine learning models for predicting startup success.

9. Research Plan

9.1. Gantt Chart



Predicting Startup Success by doing an IPO or being Acquired using Machine Learning Models

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

Period Highlight: # ▨ Plan Duration ▨ Actual Start ■ % Complete ▨ Actual (beyond plan) ■ % Complete (beyond plar

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Research Topic Submission	1	3	1	3	100%
Research Proposal Submission	4	6	4	4	100%
Literature Search	10	3	8	2	100%
Literature Review	11	3	10	2	25%
Methodology: Data Description	13	1	0	0	0%
Methodology: Data Preparation	13	2	0	0	0%
Methodology: Modeling & Inference	14	3	0	0	0%
Methodology: Evaluation	15	2	0	0	0%
Buffer: Mid Thesis Submission	17	1	0	0	0%
Experimentation: Data Preparation	18	2	0	0	0%
Experimentation: Modeling & Inference	19	3	0	0	0%
Experimentation: Evaluation	20	2	0	0	0%
Results & Discussion	22	1	0	0	0%
Conclusion	23	1	0	0	0%
Buffer: Final Thesis Submission	24	4	0	0	0%

Figure 9.1.1

Note:1 Calendar Week = 1 Period

9.2 Risk Mitigation and Backup Plan

The possible risks that could occur during the thesis completion and their respective backup plans are outlined below:

Table 9.2.1

Risk Mitigation	**Backup Plan**
Due to personal challenges, health concerns, or professional obligations, the candidate's ability to perform research is hindered, impacting the project schedule.	Candidates need to make sure that they have buffer time in research management. If candidates want an extension, then they need to inform the University administration or Upgrade.
Speciale hardware such as GPUs is unavailability for use.	Candidates can use cloud-hosted GPUs.

References

Klačmer Čalopa, M., Horvat, J. and Lalić, M., (2014) Analysis of financing sources for start-up companies. Management: journal of contemporary management issues, 19(2), pp.19-44. Available At : https://hrcak.srce.hr/file/196722 [Accessed 20 April 2024]Lyu, S., Ling, S., Guo, K., Zhang, H., Zhang, K., Hong, S., Ke, Q. and Gu, J., 2021. Graph neural network based VC investment success prediction. arXiv preprint arXiv:2105.11537.

Lyu, S., Ling, S., Guo, K., Zhang, H., Zhang, K., Hong, S., Ke, Q. and Gu, J., (2021) Graph neural network based VC investment success prediction. *arXiv preprint arXiv:2105.11537*. Available At : https://arxiv.org/pdf/2105.11537 [Accessed 20 April 2024]

Bento, F.R.D.S.R., (2018) Predicting start-up success with machine learning. *Universidade Nova de Lisboa*. Available At: https://run.unl.pt/bitstream/10362/33785/1/TGI0132.pdf [Accessed 20 April 2024]

Arroyo, J., Corea, F., Jimenez-Diaz, G. and Recio-Garcia, J.A., (2019) Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, *7*, pp.124233-124243. Available At : https://ieeexplore.ieee.org/abstract/document/8821312/ [Accessed 20 April 2024]

Pasayat, A.K. and Bhowmick, B., (2021) An evolutionary algorithm-based framework for determining crucial features contributing to the success of a start-up. In *2021 IEEE Technology & Engineering Management Conference-Europe (TEMSCON-EUR)* (pp. 1-6). IEEE. Available At : https://ieeexplore.ieee.org/document/9488587 [Accessed 19 April 2024]

Attygalle, T.I., Withanaarachchi, A.S. and Jayalal, S., (2023) Factors Influencing the Success of Software Startups in Sri Lanka: A Comparative Analysis using SmartPLS & SEMinR. In *2023 International Research Conference on Smart Computing and Systems Engineering (SCSE)* (Vol. 6, pp. 1-8). IEEE. Available At: https://ieeexplore.ieee.org/document/10215016 [Accessed 21 April 2024]

Dellermann, D., Lipusch, N., Ebel, P., Popp, K.M. and Leimeister, J.M., (2021) Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. *arXiv preprint arXiv:2105.03360*. Available At: https://arxiv.org/abs/2105.03360 [Accessed 22 April 2024]

Krishna, A., Agrawal, A. and Choudhary, A., (2016) Predicting the outcome of startups: less failure, more success. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798-805). IEEE. Available At: https://ieeexplore.ieee.org/document/7836749 [Accessed 19 April 2024]

Misra, A.K., Jat, D.S. and Mishra, D.K., (2023) Startup Success and Failure Prediction Algorithm Using k-Means Clustering and Artificial Neural Network. In *2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)* (pp. 190-195). IEEE. Available At: https://ieeexplore.ieee.org/abstract/document/10284936/ [Accessed 19 April 2024]

Kim, J.Y. and Park, H.D., (2017) Two faces of early corporate venture capital funding: Promoting innovation and inhibiting IPOs. *Strategy Science*, *2*(3), pp.161-175. Available At: https://pubsonline.informs.org/doi/abs/10.1287/stsc.2017.0032[Accessed 22 April 2024]

Al Rahma, Y.H. and Al-Alawi, A., (2023) How to Evaluate Success Startups: Case of FinTech and Cybersecurity in the GCC Venture Capital Market. In *2023 International Conference On Cyber Management And Engineering (CyMaEn)* (pp. 469-473). IEEE. Available At: https://ieeexplore.ieee.org/document/10051035 [Accessed 25 April 2024]

Sharchilev, B., Roizner, M., Rumyantsev, A., Ozornin, D., Serdyukov, P. and de Rijke, M., (2018) Web-based startup success prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management* (pp. 2283-2291). Available At: https://dl.acm.org/doi/abs/10.1145/3269206.3272011 [Accessed 23 April 2024]

Żbikowski, K. and Antosiuk, P., (2021) A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), p.102555. Available At: https://www.sciencedirect.com/science/article/pii/S0306457321000595 [Accessed 21 April 2024]

Yin, D., Li, J. and Wu, G., (2021) Solving the data sparsity problem in predicting the success of the startups with machine learning methods. *arXiv preprint arXiv:2112.07985*. Available At: https://arxiv.org/abs/2112.07985 [Accessed 23 April 2024]

Tomy, S. and Pardede, E., (2018) From uncertainties to successful start ups: A data analytic approach to predict success in technological entrepreneurship. *Sustainability*, *10*(3), p.602. Available At: https://www.mdpi.com/2071-1050/10/3/602 [Accessed 28 April 2024]

Pasayat, A.K., Bhowmick, B. and Roy, R., (2020) Factors responsible for the success of a start-up: A meta-analytic approach. *IEEE Transactions on Engineering Management*, *70*(1), pp.342-352.- Available At: https://ieeexplore.ieee.org/document/9184880[Accessed 28 April 2024]